

FILE COPY

2

## A RAND NOTE

AD-A217 309

Middle-Term Loss Prediction Models for the  
Air Force's Enlisted Force Management System:  
Information for Updating

Michael P. Murray

December 1989

DTIC  
ELECTE  
JAN 30 1990  
S E D

DISTRIBUTION STATEMENT A

Approved for public release;  
Distribution Unlimited

RAND

90 01 30 062

The research reported here was sponsored by the United States Air Force under Contract F49620-86-C-0008. Further information may be obtained from the Long Range Planning and Doctrine Division, Directorate of Plans, Hq USAF.

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of The RAND Corporation do not necessarily reflect the opinions or policies of the sponsors of RAND research.

# A RAND NOTE

N-2764-AF

Middle-Term Loss Prediction Models for the  
Air Force's Enlisted Force Management System:  
Information for Updating

Michael P. Murray

December 1989

Prepared for the  
United States Air Force



Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-11	

RAND

## PREFACE

This Note describes procedures for updating the middle-term loss equations that will be used in the Air Force's Enlisted Force Management System (EFMS). The EFMS will include a variety of loss models, distinguished by the time horizon of their prediction (short or middle term) and whether such predictions are disaggregated by occupational specialty.

This document concerns the middle-term aggregate and disaggregate loss models. Their predictions are designed to be most accurate between one and six years into the future. For an overview of the EFMS, see Grace M. Carter, Jan M. Chaiken, Michael P. Murray, and Warren E. Walker, *Conceptual Design of an Enlisted Force Management System for the Air Force*, The RAND Corporation, N-2005-AF, August 1983. Initial specifications for the middle-term loss models and details of their estimation are given in Grace Carter, Michael Murray, R. Yilmaz Arguden, Marygail Brauner, Allan Abrahamse, Harvey Greenberg, and Deborah Skoller, *Middle-Term Loss Prediction Models for the Air Force's Enlisted Force Management System: Specification and Estimation*, The RAND Corporation, R-3482-AF, December 1987.

Updating the models involves four activities:

1. adding data to the files used to estimate the equations
2. reestimating the existing specifications of the equations
3. exploring possible respecifications of the equations to exploit the additional data or to accommodate new EFMS needs
4. testing and evaluating new versions of the equations

This Note concentrates on the first three activities. The fourth is treated in Allan F. Abrahamse, *Middle-Term Disaggregate Loss Model Test and Evaluation: Description and Results*, The RAND Corporation, N-2688-AF, May 1988.

The Note should be of most interest to the Air Force analysts who are charged with maintaining and updating the EFMS. More generally, it should be of interest to modelers who have to deal with the problem of keeping their models up to date.

The work described here is part of the Enlisted Force Management Project (EFMP), a joint effort of the Air Force (through the Deputy Chief of Staff for Personnel) and The RAND Corporation. RAND's work falls within the Resource Management Program of Project AIR FORCE. The EFMP is part of a larger body of work in that program concerned with the effective utilization of human resources in the Air Force.

## ACKNOWLEDGMENTS

I am especially indebted to Marygail Brauner. Her imaginative and painstaking work in debugging Year-At-Risk (YAR) files set a standard that will be a lofty goal for analysts who create future YAR files. The diagnostic procedures for updating YAR files described in this Note are drawn from her work. I am also grateful to Dan Relles who taught me much about making the YAR FORTRAN code intelligible to people besides myself, to Glenn Clemens for allowing me to participate in his initial updating exercise, to James Hodges for an insightful and substantive review of an earlier version, and to Warren Walker who encouraged me to create this updating document.

## CONTENTS

PREFACE .....	iii
ACKNOWLEDGMENTS .....	v
FIGURES AND TABLES .....	ix
ACRONYMS .....	xi
Section	
I. INTRODUCTION .....	1
II. UPDATING DATA FILES .....	3
Process .....	3
Perspectives .....	8
III. UPDATING THE LOSS EQUATIONS .....	12
Process .....	12
Perspectives .....	13
Appendix	
A. UPDATING THE FIRST TERM ETS MIDDLE-TERM LOSS PREDICTION MODEL .....	35
B. ABBREVIATIONS FOR THE EQUATIONS .....	57
REFERENCES .....	59

## FIGURES

A.1.	Plot of unemployment variable versus residuals from first order autoregressive time trend model of Table A.9. . . . .	50
A.2.	Plot of military/civilian wage ratio variable versus residuals from first-order autoregressive time trend model of Table A.9 . . . . .	51

## TABLES

1.	Non-AFSC variables whose frequencies should be checked when updating the YAR file . . . . .	5
2.	AFSC variables whose frequencies should be checked when updating the YAR file . . . .	6
3.	Cross tabs that should be computed when updating the YAR file. . . . .	6
4.	The decision groups with middle-term loss equations . . . . .	12
5.	Demographics in the loss equations . . . . .	21
6.	Airmen's circumstances in the loss equations. . . . .	22
A.1.	Comparison of original specification of first term ETS loss model using data from YAR 2.75 and YAR 3.0 . . . . .	37
A.2.	Original specification of first term ETS loss model excluding and including early reenlistments . . . . .	39
A.3.	Original specification of first term ETS loss model excluding and including cases missing WBONC . . . . .	40
A.4.	Original specification of first term ETS loss model excluding and including six-year enlistees who enlisted before July 1974. . . . .	41
A.5.	Original specification of first term ETS loss model for original sample period (7607-8306) and enlarged sample period (7607-8505). . . . .	42
A.6.	Original specification of first term ETS loss model with fits for full sample period and for three subperiods . . . . .	43
A.7.	Original specification of first term ETS loss model with and without a time trend. . . . .	44
A.8.	Original specification of first term ETS loss model controlling for ultimate AFSC and the modified occupational code . . . . .	45
A.9.	Original specification of first term ETS loss model correcting for first-order serial correlation. . . . .	48
A.10.	Modified Box-Pierce $\chi^2$ test statistics for remaining serial correlation at up to 12, 24, 36, and 48 months after correcting for first-order serial correlation among the residuals of the autoregressive models . . . . .	51
A.11.	Updated specification of first term ETS loss model . . . . .	53

## ACRONYMS

AFSC	Air Force Specialty Code.
ARIMA	Autoregressive Integrated Moving Average.
CONUS	Continental United States.
DoD	Department of Defense.
EAGL	Enriched Airman Gain/Loss (file).
EFMP	Enlisted Force Management Project.
EFMS	Enlisted Force Management System.
ETS	Expiration of Term of Service.
OETS	Original Expiration of Term of Service.
OLS	Ordinary Least Squares.
SAS	Statistical Analysis System.
TAFMS	Total Active Federal Military Service.
YAR	Year At Risk (file).

## I. INTRODUCTION

This Note describes procedures for updating the middle-term loss equations that are used in the Enlisted Force Management System (EFMS).

Updating involves four activities:

1. adding data to the files used to estimate the loss equations;
2. reestimating the current specifications of the equations;
3. exploring possible respecifications of the equations to exploit the additional data or to accommodate new EFMS needs; and
4. testing and evaluating the new versions of the equations intended for use in the EFMS.

Adding data to the files used for estimating the models requires understanding the structures of three data files: (1) the Enriched Airman Gain/Loss (EAGL) file, (2) the Year-At-Risk (YAR) file, and (3) the analysis files drawn from the YAR file for use as direct inputs into the estimation programs. Adding data also requires understanding the programs that create the YAR and the analysis files.

Reestimating the current specifications of the loss equations only requires understanding the programs that calculate the estimates. Exploring possible respecifications is more demanding. It requires understanding: (1) the statistical strategy underlying the estimation procedures, (2) the perils for estimation inherent in the available data, (3) the uses to which the loss equations will be put, (4) the programs for calculating estimates, and (5) how to adapt the equations in response to information from the testing and evaluation exercise.

Testing and evaluating the new versions of the loss equations requires understanding: (1) the testing programs, (2) the performance criteria used to evaluate the performance of the loss equations, (3) the purposes to which the loss equations will be put, and (4) the "blending" process by which loss estimates for individuals in a given year at risk are transformed into estimates of loss rates for the Air Force in a given fiscal year.

This Note is not intended to replace the existing technical documents that explain the various data files (Brauner et al., 1989; Murray et al., 1989), the estimation of the current version of the loss equations (Carter et al., 1987), or the test and evaluation process (Abrahamse, 1988). Rather, its purpose is to guide the analysts who will update the middle-term loss equations, a process that will require integrating the contents of this and the other documents. Consequently, the emphasis here is on broad descriptions of tasks and their interrelationships, on motivations for approaches to tasks, and on opportunities and pitfalls that the analysts are likely to encounter, rather than on technical detail.

The following sections treat updating the data files and the reestimation and respecification of the equations. Testing and evaluation is discussed in detail by Abrahamse (1988).

## II. UPDATING DATA FILES

### PROCESS

The Enriched Airman Gain/Loss (EAGL) file (see Brauner et al., 1989) provides the longitudinal information on enlisted personnel necessary for estimating loss equations. These data are combined with supplemental historical data on military and civilian compensation, civilian unemployment rates, bonuses available to airmen, and Air Force Specialty Code (AFSC) conversions (see Walker and McGary, 1989) to form the Year-At-Risk (YAR) file (see Murray et al., 1989), from which samples are drawn for estimating and validating the middle-term loss equations.

The units of observation in the YAR file are called "years at risk." Associated with each term an airman serves are one or more years at risk. The first year at risk in an airman's term of service is the twelve-month period beginning with the calendar month in which the term itself begins. The second year at risk in an airman's term is the twelve months following the first year at risk. Subsequent years of risk cover subsequent twelve-month periods. The last year at risk in a given term of service is the year at risk covering the date on which the airman ends that particular term by either reenlisting or leaving the service.

The middle-term loss equations estimate the probability that an airman will leave the service in a given year at risk of a given term of service. For reenlistment-eligible airmen who do not leave the service, the equations also estimate the probability that an airman will extend rather than reenlist in a given year at risk.

Updating the EAGL and supplementary historical data files to add years of data is a separate task from updating the loss models; it is discussed no further.

Updating the YAR file is the first step in updating the middle-term loss equations. The step uses:

1. the updated EAGL file,
2. updated supplementary historical data files,
3. the YAR FORTRAN subroutines,

4. a system code that links the data and the FORTRAN code and executes the FORTRAN program, and
5. a SAS program that creates a SAS data set containing the YAR file.

The FORTRAN code that creates the YAR assumes a specific format for the EAGL data. Consequently, any revisions in the structure of the EAGL file will necessitate changes in the YAR FORTRAN code.

Each time the YAR file is updated, the maximum number of records for an airman grows. Consequently, the parameters in the FORTRAN code that define array lengths must be changed. The maximum possible numbers of snapshots, years at risk, and transactions must be set to accommodate the length of the YAR file used. Suppose one wishes to set the maximum number of snapshots to 17, the maximum number of years at risk to 31, and the maximum number of transactions to 47. The following global change in the YAR FORTRAN code accomplishes this:

```
change          "parameter (j1=14,j2=28,j3=42)"
                to
                "parameter (j1=17,j2=31,j3=47)"
```

Across terms, years at risk can overlap. For example, an airman whose first term began in June 1976 and whose fourth year at risk for that term ends in June 1980 may reenlist in, say, March 1980; the last year at risk of the first term is 7906-8006, and the first year at risk of the second term is 8003-8103. For a detailed description of the year at risk concept, see Murray et al., 1989.

After a new YAR has been created, it is necessary to look for errors that were made in creating it. If errors are found that can be corrected, the corrections must be made and the updated file re-created. Looking for errors in new files is a tedious job that requires considerable patience and painstaking attention to detail on the part of the data diagnostician. The process is more an art than a science; it has no fixed rules, but some guidance is provided below. What follows assumes that the reader is familiar with the variables in the YAR file (see Murray et al., 1989).

The first step is to look at the frequencies of some of the variables to find if any of their values are out of their expected range. Table 1 contains a list of suggested variables to look at.

Table 1

NON-AFSC VARIABLES WHOSE FREQUENCIES SHOULD BE  
CHECKED WHEN UPDATING THE YAR FILE

BDEPEN	YFMS12	TMBMLT	XXTDOS1
BIAGE	YGRD12	TPRVBM	XGRADE
BIDOE	YLGP12	TNEXT	XTDCSA
BIEDUC	YMCWRAT	TNTS	XTOE
BITOE	YNTYR	TNTYR	GOODTYR
BMALE	YNXTLS	TNWBON	GOODSNAP
BMARRY	YRISKB	TOLBP	TOTSPD
BPACE	YTRMYR	TOLCP	
BRACE	YT1BP	TOLMP	
BSEX	YT1CP	TOLUEM	
SDOS	YT1MP	TPSA	
SEDU	YT1UEM	TQWBAK	
SGRD	YYOS12	TQWLST	
STAF	TABMLT	TRMDUM	
STIG	TBDOBS	TT1UEM	
SYR	TBEGIN	TXOETS	
YBMLT6	TBEND	TXTOE	
YCONUS	TBML12	TYOS12	
YDOSIN	TBRECD	XCOH	
YDOSOU	TENDLS	XCTRN	
YE1BP	TFMS12	XDOS	
YE1CPG	TGRD12	XEXTED1	
YE1MP	TLASYR	XTRNID	

If SAS is used to obtain these frequencies, be aware that (depending on the installation) PROC FREQ limits the number of values a variable can assume. Thus, for example, SAS cannot be used to find the frequencies on any of the AFSC variables. Software in addition to SAS will be needed for this diagnostic work, because frequencies on AFSC variables should be computed. Table 2 contains a list of the AFSC variables that should be checked.

Cross tabs should also be calculated so that anomalies can be identified. Table 3 lists several cross tabs that should be reviewed. They are mainly the checks for consistency. For example, the variable GOODSNAPE will be a rough indicator of the force size in a given year, so the frequency for one 10 percent sample should be approximately  $(.1) \times (\text{force size})$ . Also, if  $SYR < 0$  then GOODSNAPE should = 0. Every record with TRMDUM=1 should have TENDLS=1 unless the year at risk is the last in the file, in which case TRMDUM may equal -9; GOODTYR should be 0 when TBDOBS=1 or when TRMDUM=1.

The diagnostician needs to write an edit program that reflects the information in Murray et al., 1989. The program would look for records that do not match the descriptions of the variables. For example, if (TXOETS = -9 and TBEGIN = -9 and TBDOBS ne 1) something is wrong. Among the logical checks should be:

YRISKB < TBEND when TBEND > 0.  
 TBEGIN <= TBEND when TBEND > 0.  
 TXOETS <= YDOSOU  
 TXOETS <= YDOSIN  
 TBEGIN <= YRISKB  
 TENDLS ne -9 when YDOSOU <= June of last SYR

Table 2

AFSC VARIABLES WHOSE FREQUENCIES  
 SHOULD BE CHECKED WHEN  
 UPDATING THE YAR FILE

SDAF	TUCAF
SPAF	TUDF12
YDAF12	XCAFSC
YPAF12	XEXTAF1
YUAF12	XEXTAF2
TLAFSC	XEXTED1
TUAFPR	XEXTED2
TUAF12	

Table 3

CROSS TABS THAT SHOULD BE COMPUTED  
 WHEN UPDATING THE YAR FILE

GOODSNAP	AND	SYR
TENDLS	AND	TRMDUM
GOODYR	AND	TBDOBS
GOODYR	AND	TRMDUM
TMBLMT	AND	year of TXOETS

Special care should be taken with the bonus variables. These variables are created by a complex process that includes not only recording the bonuses by applicable date and

eligible AFSC, but also tracking AFSC codes over time. In the past there have been many errors in the bonus variables. To catch these errors one must check that, in the aggregate, the correct ranges of bonuses are offered each year and then that, for specific important AFSCs, correct bonuses are recorded for each year. This very important activity will take much time and careful thought on the part of the person checking the data.

The above checks will be certain to reveal some problems. At this point, all records from both the EAGL and YAR file for offending cases must be reviewed. For example, if (TBEGIN>TBEND) is true for some YAR records, find all the YAR and EAGL records for such individuals. *An important diagnostic tool for the updated YAR file is a program that extracts specific YAR records and corresponding EAGL records.*

Usually, bad YAR records are the result of bad EAGL data. But if the EAGL data appear to be correct, two steps should be taken, one by the YAR diagnostician who is checking the new data, the other by the YAR programmer who is responsible for maintaining the YAR FORTRAN code.

First, the diagnostician should compare all of the YAR records that share the anomaly to identify what other traits they share. It would be invaluable for the YAR programmer to know, for instance, that TBEGIN is greater than TBEND only for airmen already in the service at the start of the YAR file's sample period.

Second, the YAR programmer should trace through the logic of the YAR code to find where the values of the variables are set and exactly which EAGL variables are used for the values of the bad YAR variables. If this does not enable the programmer to identify the problem, a trace should be made of the creation of the YAR variables for the troublesome cases; the "debug" feature of FORTRAN produces such a trace.

The YAR code should be changed only when it has been unquestionably determined that the YAR records do not correctly record GOOD data from the EAGL file. Be extremely cautious about changing YAR code. The ripple effects of one small change can be enormous. To minimize the chances for error, a sample of EAGL data should be available for testing changes in the YAR code. Careful comparisons of the outputs from the YAR code before and after code changes should be made to insure that changes in the YAR code change only the troublesome cases and no others.

An efficient way to update the YAR file would be to alter only the YAR records of airmen whose EAGL records have changed. When the EAGL is updated, a file of Social

Security numbers for the airmen with new (completely new or merely revised) EAGL records can be created. YAR records in the old file should then be divided into two groups, airmen with revised EAGL records and airmen with unchanged EAGL records. Passing the new EAGL records through the YAR FORTRAN program will create new YAR records for both the airmen with completely new EAGL records and the airmen with revised EAGL records. These new YAR records should then be merged with the old YAR records for airmen whose EAGL records remained unchanged. As the YAR file grows over time, the savings in processing from this more efficient updating procedure will become considerable.

After an acceptable updated YAR file has been created, there remains one last task: The YAR flat file must be put into a SAS data set. A SAS program that does this is given in Murray et al., 1989, App. C.

## **PERSPECTIVES**

### **The YAR File and the EAGL**

Since the YAR file is derived from the EAGL, an understanding of the EAGL's structure is necessary to correctly update the YAR.

The EAGL file contains extensive information about all airmen who were in the Air Force any time during the file's historical span. The data for each airman are drawn, ultimately, from transaction records and from Uniform Airman Records (UARs). Occasionally, the source data are miscoded or incomplete.

Since the EAGL records are constructed with little manipulation of the input data, errors in the input data seldom cause errors in the EAGL records beyond the replication of the miscoded or missing value. However, the YAR files are constructed with extensive manipulation of the EAGL data. Consequently, erroneous or incomplete input data can easily influence the structure of records as well as lead to erroneous or incomplete data items.

For example, a miscoded first reenlistment transaction that appears as an unrecognizable transaction type will either appear as such in the EAGL file or, perhaps, be lost altogether. Thus, the error in the EAGL file is no more extensive than that in the source data. But such a problem in the source data and EAGL file can wreak havoc in the YAR file.

Without a reenlistment transaction to mark the end of the airman's first term, the YAR file will declare all of the airman's second term activities to have occurred in the

airman's first term. Moreover, the erroneous first term structure threatens to contaminate the airman's entire history, since a third term (a career term) could be mistakenly categorized as a second term.

The sensitivity of the YAR file structure may be reduced to errors in the individual data elements of the EAGL file by exploiting redundancies in the EAGL data when creating the YAR file. For example, in the above case, the first term is unavoidably distorted; without a first reenlistment transaction, one cannot know that the term ended when it did. But by checking two other codes, the Enlistment Eligibility Category and the Total Active Federal Military Service (TAFMS), at the supposed end of the first term, one can infer that the next term is indeed a career term, not a second term. Thus, although two terms in the airman's history are lost, the third and subsequent terms can be salvaged. Much of the current YAR code is devoted to using redundancies in the EAGL data to reduce to a very low level the proportion of YAR records that are distorted by errors in the EAGL file.

Exploiting redundancies in the EAGL file is not always possible, nor is it costless. Analysts updating the YAR file must be cautious about modifying the YAR code to accommodate additional flaws in the input EAGL data. The YAR program is complex; altering the code in one section always risks introducing subtle bugs that do more harm than the intended fix undoes. The relevant rule recommended for changes in the YAR code is:

*The YAR code should always use accurate input data correctly. If it does not, the code should be fixed. However, the YAR code should be expected to create erroneous records when given erroneous input data. Only when the erroneous YAR records are large enough in number and poor enough in quality to threaten the integrity of the estimations should the YAR code be altered to overcome the errors in the input data. In such extreme cases, alternatives that would improve the quality of the input data should be considered first. Analysts using the data should be notified if ways to detect bad records can be devised.*

Improving the quality of the input data can take several forms. The most important are identifying alternative sources for the data to use in constructing the EAGL file and improving incentives for the primary data gatherers to properly record the data. For example, in the original EAGL files, the CONUS/non-CONUS indicators for airmen were drawn directly from the variables that purported to be CONUS/non-CONUS indicators. Unfortunately, these indicators were frequently misleading. An airman who had returned to CONUS for discharge would be listed as CONUS even though the airman's most recent duty assignment had been non-CONUS. Now CONUS/non-CONUS status is derived from the airman's unit code and a specially prepared file mapping dates and unit codes to CONUS/non-CONUS assignment.

Modifying the YAR code requires an able FORTRAN programmer. Checking the modified output for inadvertently introduced errors requires an especially patient programmer, one willing to engage in exceedingly tedious checking and cross checking of processed records. The less able and patient the available programmers, the more reluctant one should be to make changes in the YAR code to accommodate errors in the input data. Personnel turnover in the Air Force will inevitably bring fluctuations in the skills and traits of the programmers available for working with the YAR, and this should be taken into account when deciding whether to modify the YAR code.

### **The Length of the YAR and the Quality of Records**

The YAR file contains records for all years at risk that began after June 1971. The documentation for the YAR file recommends, however, that analysts should not use years at risk that began before July 1976, because those records frequently are missing data. Furthermore, older records are more likely to reflect decisions made by draftees and decisions made in response to wartime conditions or experiences. Loss models for the peacetime all-volunteer Air Force should rely as little as possible on data from earlier periods.

Nonetheless, *all data from July 1971 onward should be retained in updated YAR files, which requires that updated EAGL files also retain all these data.* The consistency checks that improve the quality of YAR records often involve past data about an airman. (This is especially true for airmen whose initial enlistment did not occur within the years covered by the files.) If data from an airman's recent past are not available, the quality of current records suffers. As a consequence, data for the first few years in the YAR file, no matter which years they may be, will always be of lower quality than data for

subsequent years. Including the last years influenced by draft and war effects as the first years in the YAR file ensures that the data from the all-volunteer period are of as high quality as possible.

### III. UPDATING THE LOSS EQUATIONS

#### PROCESS

The middle-term loss equations predict losses for the ten groups of decisionmakers listed in Table 4. Some decision groups require several equations; for example, ETS decision groups require both loss and extend-given-stay equations. In all, 18 loss equations require periodic updating. These loss equations are used in the middle-term inventory projection models of the EFMS.

Updating each equation involves the same two tasks:

1. selecting and transforming an appropriate sample of years at risk and
2. estimating parameters of the equations with the new data.

Table 4

#### THE DECISION GROUPS WITH MIDDLE-TERM LOSS EQUATIONS

Decision Group	Number of Equations <sup>a</sup>
First term attrition	3
Second term attrition	1
Career term attrition	1
First term ETS	2
Second term ETS	2
Career term ETS	2
First term extension	2
Second term extension	2
Career term extension	2
Retirement	1

<sup>a</sup>The names and abbreviations for the equations are given in App. B.

A set of SAS programs that perform these tasks must be developed. To select and transform the data the programs must:

- select good years at risk for a specified sample period,
- restrict cases to the decision group in question (which sometimes requires careful definition),
- redefine missing value indicators and delete some cases with values for particular variables missing,
- create transformed variables as needed, and
- add any special purpose data not contained in the YAR records.

*Equation parameters are generally estimated in two steps:*

1. PROC ABSORB in SAS is used to estimate the parameters of all variables except the AFSC dummies, and
2. a special purpose SAS program is used to estimate the parameters for AFSC dummies conditional on the other parameter estimates.

This two-step procedure is essentially equivalent to ordinary least squares (OLS) but avoids the computational problems that can arise when inverting a matrix that is larger than 300\*300 (as the more than 300 AFSCs would require in OLS estimation). Some of the models require slight adaptations of the two-step procedure. For example, the models of losses from the career force are estimated using the SAS procedure PROC REG (Carter et al., 1987).

Between the selection of data and the estimation of parameters lies the most demanding task in updating the loss equations—deciding on the specifications of the equations to be estimated. This task requires repeated application of the estimation programs to try and to reject alternative candidate specifications until an acceptable version is found. After the chosen specification is subjected to testing and evaluation with additional data, further respecification may be required.

## **PERSPECTIVES**

To update the middle-term loss equations, one must have a clear understanding of (1) the modeling strategy suitable for these equations, (2) the structure of the equations, (3) the variables that have been found important in the past and those most likely to prove important in the future, (4) the estimation techniques suitable for the equations, (5)

any new policy uses to which the EFMS may be put, and (6) what prospects additional data hold for improved modeling.

### **The Modeling Strategy**

The middle-term loss equations are used to forecast airman losses one to six years in the future, particularly how changes in the economic opportunities of airmen, measured by unemployment rates and military compensation (most importantly bonuses), affect losses. Forecasting accuracy determines the choice among alternative models of loss behavior.

Analysts trained in economics must discipline themselves not to approach the loss equations in terms of their power to explain airman behavior. Accurate forecasting is the primary goal of the equations; explanation is only a secondary goal.

The short-term loss modeling exercise offers a useful lesson in this regard. Serial correlation in airman loss rates is very strong from one month to the next. Consequently, a pure time series model of airman losses yields much better forecasts from month to month than does a behavioral economic model. For the EFMS, a time series model is definitely preferable to a behavioral model for short-term analysis, even though the former offers no explanation for the loss behavior. A model that incorporates both time series and behavioral information may perform little better than a simpler, pure time series model; in such a case the added complexity of the mixed model is not worthwhile.

Loss rates are much less correlated from year to year than from month to month. Consequently, pure time series models would not be useful for the middle-term analysis. In the middle term, the choice is among behavioral models of varying complexity. A simple model is chosen that posits the probabilistic outcome from any one decision to be a function of the airman's traits, circumstances, and economic opportunities. In particular, more sophisticated models have been avoided that recognize the interdependence among an airman's choices at different times.

The most sophisticated behavioral model for examining airman losses is that of Gotz and McCall (1984). It offers a consistent framework for explaining how complicated changes in airman compensation, such as changes in the retirement system, would alter loss decisions by airmen. However, the model's complexity precludes incorporating even small numbers of airman traits (such as race, sex, marital status, AFSC, and AFQT scores) into an estimated model. Consequently, predictions of losses using such a model would not be able to capture the systematic variation in loss rates

across demographic groups or across AFSCs. The simpler specifications chosen for the middle-term equations are able to incorporate these covariates; consequently, their forecasting performance will generally be better than that of the more sophisticated model.

The EFMS includes a Gotz-McCall style loss model, which Arguden (1986) developed. His analysis emphasizes that simpler specifications can work effectively across a wide variety of changes in economic opportunities. He also clearly identifies the kinds of compensation changes whose effects simpler models will not forecast well. For example, the effects of military pay increases can be forecast adequately by simple models, but the effects of a new military retirement system cannot.

The new retirement system that went into effect in 1986 illustrates an important problem for users and maintainers of the middle-term loss equations. How should the problem identified by Arguden be overcome? Should the simpler modeling approach be abandoned? Should the problem be ignored? The answer to each of these questions is no.

The flexibility of the simple models that allows them to incorporate many covariates is too valuable to abandon. Policy relevant variables, such as AFSC, and behaviorally influential variables, such as unemployment rates, are too important to ignore. But the potential forecasting errors that the new retirement system might cause are also too great to ignore. Two things should be done:

1. When sufficient data are available to estimate the difference in loss rates between the new and old regime, an updated model should incorporate a variable differentiating between the two retirement system periods; and
2. Until such data are available, the Arguden model should be used to estimate the average forecast error that is likely to occur from using the simpler, misspecified equations.<sup>1</sup>

---

<sup>1</sup>Such errors are likely to be small for many years. Airmen already in the service face the old retirement system rules. Since the new rules are likely to alter loss rates markedly only for second term and career airmen, it will be ten or twelve years before the new rules affect forecasts much. In the early years of the EFMS, no such adjustments to forecasts will be necessary.

The example of the retirement system change highlights two important features of the middle-term loss equations in the EFMS.

First, *repeated updating of the loss equations as new data become available is necessary to reduce the contamination of forecasts by specification errors stemming from changes in the underlying environment in which airmen make their decisions.* In the extreme, drastic changes can be adjusted to by restricting the sample data used to the period of a new regime, a step that would assume all previous experience is irrelevant to the new circumstances. (In essence, this is what happens when data affected by wartime and draft-time experiences are not used.)

Second, the inclusion of Arguden's model as a part of the EFMS provides an independent check on the validity of the middle-term loss equations in the face of extensive adjustments to the military compensation system. *Users of the middle-term loss equations should periodically use Arguden's model to see if adjustments to forecasts from the simpler models are obviously warranted.* Such checks should be made whenever there are complicated changes in military compensation. The analysis should ascertain which decision groups will be strongly affected by the compensation changes, and when.

Models more complex than the middle-term loss equations but less complex than the Gotz-McCall model have been used in several services for analyzing military compensation policies. (They are called ACOL and PVMOL models.) These models are simpler to estimate than the Gotz-McCall model, but, as Arguden (1986) shows, they perform poorly in forecasting the effects of some complex changes in military compensation. Consequently, these models would need to be supplemented by a model such as Arguden's, just as the simple equations need to be.

These intermediate models would perform comparably to the simpler models for the most frequent changes in economic opportunities, but they would be computationally more demanding. Their only advantage would be that for some complex compensation changes, their forecasts would reflect some of the more subtle behavioral considerations that the Arguden model would. But since the Arguden model is needed to check both the intermediate and simple models in the face of complex compensation changes, this advantage is of little importance for the EFMS.

## **The Structure of the Models**

The middle-term loss equations fall into four groups:

1. first and second term attrition and ETS loss models,
2. career attrition and ETS loss models,
3. the retirement loss model, and
4. extension loss models.

All the equations are estimated as linear probability models in which the probability of loss (or extension) is a linear function of a set of explanatory variables. Linear probability models have the undesirable characteristic that calculated probabilities can be greater than one or less than zero. To overcome this deficiency, when the loss equations are used in the middle-term inventory projection models of the EFMS, estimated probabilities outside the required range are made zero or one.

The seeming sameness in the structures of the various models masks important differences among them. In updating the equations, careful attention should be given to these differences.

### **First and Second Term Attrition and ETS Loss Models**

These are the most straightforward models. The attrition models treat losses in years at risk that begin more than 12 months before an airman's originally scheduled expiration of term of service (OETS). In the attrition models, the probability of loss for a given time period in the term is the dependent variable. The ETS models treat the outcomes during the year at risk beginning 12 months before an airman's OETS. (Early release program losses and early reenlistments are included in the ETS models.) In the ETS models, either the probability of loss or the probability of extending given staying is the dependent variable.

In the first term attrition model, the time periods used are (1) the first two months of service, (2) the remainder of the first year at risk, and (3) subsequent years at risk. In the second term attrition model, all years at risk are treated in a single equation.

The three first term time periods correspond in essence to basic training, initial specialty training, and post training periods. These distinctions are likely to be needed in updated specifications. Currently, differences across years at risk in first term post-training attrition and second term attrition are captured only by dummy variables that

interact with length of enlistment. Additional data might support more detailed specifications of these differences, and this possibility should be examined when updating the equations.

In the first term ETS loss and extend-given-stay models, there is a difficulty in defining the appropriate population. The Early Release program sometimes allows substantial numbers of airmen to leave the service before the contractual end of their first term. The current specification treats these early release losses like ordinary ETS losses because the data do not support a different treatment. However, since some airmen released early might have changed their minds and reenlisted had they completed their terms, it is possible that the Early Release program raises loss rates above what they would otherwise be. In updating the first term ETS equations, some attention should be given to exploring this possibility.

The second term ETS loss and extend-given-stay specifications offer a different unresolved issue. More complex models of loss behavior suggest that loss rates in the second term for a cohort of airmen who entered the service at the same time will depend on the loss rate at the end of the first term. If Air Force policies induce an above average proportion of a cohort to stay for a second term, the loss rate at the end of that second term will tend to be higher than average. The current specification of the second term loss equation does not incorporate a cohort's first term retention rate as an explanatory variable because inclusion was not supported in the current sample. But in updating the model, this effect should be looked for once more. (The inclusion of an airman's having received a first term bonus as a variable in the second term equation captures one component—and perhaps the most important component—of this effect.)

### **Career Attrition and ETS Loss Models**

The career equations differ from the first and second term equations in their attention to years of service.

In the first and second term, airmen making, say, an ETS decision will differ little in years of service. But in career terms, airmen making ETS decisions may have as few as ten or as many as 18 years of service. Interactions between years of service and other variables appear in the career equations but not in the first or second term equations. (These interactions are not likely to disappear when the equations are updated.)

Furthermore, as airmen approach 20 years of service and, thus, eligibility for retirement benefits, the incentive to stay in the service to the 20 year point grows

steadily. To reflect this, the career equations constrain estimated loss rates to shrink steadily toward zero as an airman's years of service approach 20.

These constraints on the equations' parameters require a computer routine that can perform least squares while imposing linear constraints across the parameters of a linear equation. PROC ABSORB in SAS, the routine that vastly simplifies the treatment of AFSCs in estimating the first and second term models, cannot do this. The career models must be estimated with a different procedure; PROC REG was used to estimate the current versions of the career equations.

The career equations were estimated with grouped data. Instead of using each airman's data as a separate observation, as was done to estimate the other equations, groups of airmen were formed and the loss rates for each group calculated; these groups were the observations used to estimate the equations. (The groups were defined by combinations of demographic traits and circumstances of the airman in the Air Force, such as AFSC and grade.) The dependent variables in the loss equations became the loss rates for the groups of airmen; the explanatory variables became the means of the relevant variables within each observed group.

The PROC REG routine in SAS does not require group data; individual data could have been used to estimate these equations with this procedure. Initially, grouped data were used to keep down computing costs; later, examination showed that exploiting within-group variability would not alter the parameter estimates, so the change to individual observations in estimation was not made. When the career term equations are updated, the choice between individual and grouped data should be reconsidered.

### **The Retirement Loss Model**

The retirement equation differs from the first, second, and career equations. A single equation is used for all retirements instead of having separate equations for attrition losses before ETS, ETS losses, and losses after extensions. Years-to-ETS enters the retirement equation solely as a categorical variable. Years of service is of especial importance in the retirement model.

There were several reasons for this choice. First, year of service is used in Air Force retirement policies to determine the conditions of retirement. For example, the amount of retirement benefits an airman will receive increases with the number of years served. Second, under certain circumstances, retirement-eligible airmen can leave the

service on seven days notice. Finally, the distinction between reenlistments and extensions in the retirement-eligible years is not as meaningful as in the earlier years.

Consequently, the year at risk in this model is not based on the date of OETS but is defined by the airman's year of service (based on TAFMS and the date of enlistment). Environmental variables used in this model that appear on the YAR must be recalculated from the perspective of this date instead of the year at risk on the YAR. For instance, for retirement-eligible airmen, moving averages for the unemployment rate drawn from the YAR must be reassigned to cover the retirement year at risk.

The high year of tenure rules imposed by the Air Force are the single most important influence on when eligibles retire. Another strong influence is the policy covering obligated service that results from promotion to grade E-7 or higher. If these policies ever change, or if a substantial number of waivers are granted, then the analyst will have to respecify the equation to capture the new policy during the period of its effect.

### **Extension Loss Models**

Extension loss equations are built around the ETS established when an airman reenlisted. For first, second, and career term airmen there are two extension loss equations. The first is for airmen in a year at risk beginning after their original ETS but more than 12 months before their extended ETS. The second is for airmen in a year at risk beginning after their original ETS and within 12 months of their extended ETS. The structure of this model is likely to change when updated. Frequent policy shifts in the original sample period absorbed many degrees of freedom on this model so that many alternative specifications could fit the data about equally well. With additional years of data we will be better able to distinguish among alternative specifications.

### **The Variables in the Models**

Four classes of variables appear in the middle-term loss equations. There are variables for:

- airmen's demographic traits,
- airmen's circumstances in the service,
- airmen's economic opportunities, and
- Air Force policy periods.

Not all classes appear in every model in the current version of the loss equations. Some variables may be deleted from or added to individual equations when the equations are updated. In each class, there are variables to which especial attention should be given when updating particular models.

**Airmen's Demographic Traits.** Table 5 reports the demographic variables that appear in the middle-term loss equations. As the table makes clear, demographic influences attenuate the longer an airman is in the force.

Table 5

DEMOGRAPHICS IN THE LOSS EQUATIONS<sup>a</sup>

Variable	1st Term Attrition			1st Term ETS		2nd Term ETS		ret
	1att2	1att10	1atts	1ets	1egs	2ets	2egs	
Age	x	x	x					
Age*term length		x	x					
Race	x			x		x		
Sex	x	x		x	x		x	
Marital status	x	x	x	x	x	x	x	
Dependents		x	x					
Sex*marital status				x	x			
Sex*race		x	x					
Sex*occupation			x					
Education	x	x	x		x	x	x	x
Intelligence	x	x	x		x			

<sup>a</sup>Demographics do not appear in the second term attrition equations, the career equations, or the loss from extension equations.  
Appendix B defines the abbreviations for the equations.

A part of updating the model is to look for the emergence of new demographic effects and the disappearance of old demographic effects. As the table shows, this should include interactions as well as main effects. If the proportion of women in the service grows, the extensive interactions between sex and other variables may eventually warrant separate equations for men and women. As women's roles in the service become more like men's, and as their civilian options become more alike, the observed differences in their loss rates may become less.

**Airmen's Circumstances in the Service.** Table 6 reports the variables pertaining to an airman's circumstances in the service that appear in the middle-term

Table 6

AIRMEN'S CIRCUMSTANCES IN THE LOSS EQUATIONS<sup>a</sup>

Variable	1att10	1atts	1ets	1egs	1x1nd	1xld	2att	2ets	2egs
Term of enlistment			x	x					
Grade							x	x	x
Year of term							x		
Years of service		x				x		x	x
Toe*demographics	x	x							
Toe*grade								x	
Toe*unemp.			x						
Toe*yos	x								
AFSC			x	x				x	x
Career field		x					x		
Career field group					x	x			
C.f.group*demog.		x							
	2x1nd	2xld	catt	cets	cx1nd	cxld	ret		
Term of enlistment			x		x				
Grade			x	x	x		x		
Years of service		x	x	x	x	x	x		
Toe*yos			x		x				
Career field							x		
Career field group	x	x	x	x	x	x			
C.f.group*yos					x				
Grade*yos					x				
High year of tenure							x		
Year in grade							x		
Years to OETS			x				x		

<sup>a</sup>Appendix B defines the abbreviations for the equations.

loss equations. As with demographic variables, the appearance and disappearance of influences should be checked in each updating of the equations.

Most important, as more data become available, AFSC-specific effects should be tested for in the models that currently have only career field or career field group effects.

The grade variable poses particular statistical problems because of its potential endogeneity—the possible feedback from loss rates to grade. Incorporating grade into the first term model is especially difficult, because airmen who do not leave early are more likely to be promoted in their fourth year and it is fourth year promotion that

accounts for almost all the variation in grade among first termers. Consequently, grade has an especially strong spurious correlation with staying beyond ETS in the first term. *Grade effects should not be estimated with first term data. If grade effects in later terms are found to be unbiased, however, some adjustment for grade in the first term equations based on estimates from those other terms might be attempted.*

**Airmen's Economic Opportunities.** Economic variables appear in all but the attrition equations. Unemployment appears in all nonattrition equations except the first term extend-given-stay equation. The military civilian pay ratio appears in all nonattrition equations except the retirement and first term extend-given-stay equations. Bonuses appear in the first and second term nonattrition equations.

The unemployment variable is a moving average of monthly unemployment rates. This was chosen because airmen are believed to respond more to expectations of longer term employment prospects than to short term fluctuations in unemployment rates. With a longer data series, this presumption could be empirically tested.

The initial explorations used race- and sex-specific unemployment rates but only the race- and sex-specific constants were affected by this; estimated unemployment effects were not altered. Nonspecific unemployment rates are easier to track and easier to implement in the Inventory Projection Models (IPMs), so average unemployment rates were used in the equations. Civilian unemployment patterns will probably not change enough to warrant the use of sex- and race-specific unemployment rates in the future.

The unemployment variable in the equations does not vary with the ages of the airmen. However, the unemployment rate that appears in the equations is not an average rate across all ages, it is the rate for 20- to 24-year-olds. (See Murray et al., 1989, for the definition of the unemployment rate contained in the file.) Unemployment rates across age groups are highly correlated from month to month; consequently, the empirical performance of the equations is not affected by the use of one age group throughout. If it would be easier to explain the equations to audiences for the EFMS's results with an average rate of unemployment over all age groups, that change could be made when updating the equations.

The DoD is no longer compiling military/civilian pay ratio used in the equations. The Air Force will have to construct its own series in the future. In updating the model, it will be important to see if the pay series developed for the EFMS is strictly comparable to the old DoD series.

The current version of the retirement model does not include the military/civilian wage variable. A longer time series might enable the inclusion of this variable in the retirement model.

A drawback to the military/civilian pay ratio is that it includes no variation in civilian opportunities across AFSCs. An effort was made to build AFSC-specific civilian pay variables that would provide within-period variation in military/civilian pay ratios, but it failed. First, there are currently no good measures of skill-specific civilian wages. Second, an airman's AFSC is only weakly correlated with his civilian sector job.

An extensive series of airman exit interviews might reveal strong AFSC wage opportunity differences that could be incorporated into updated loss models. However, the efficiency gains from such an exercise will grow small over time. The AFSC opportunity differences are almost surely stable over time; hence the AFSC-specific constants will capture these effects sufficiently for forecasting purposes. The only advantage of identifying wages' contribution to the AFSC-specific effect would be a more precise estimator for the effect of military pay changes.

The bonus variables used in the model are based on an airman's ETS, AFSC, and years of service. The data require care in construction and pose a formidable problem for the builders of the YAR file. For analysts updating the loss equations, the most serious worry arises from the tracking of AFSCs over time. Periodically, two or more AFSCs are combined into a new category. When two AFSCs with widely different historical loss rates are merged into a single AFSC, the estimated bonus effects in the equations can be markedly altered. *A research project that examines how AFSC changes affect estimated bonus coefficients in various subperiods of the data is a high priority item in the long term maintenance of the loss equations.* The problem's potential came to light very late in this work and has not received the attention it needs. The variable for bonus opportunities in AFSCs other than the airman's own suffers similar difficulties and should be included in the suggested research project.

Whether or not an airman received a Zone A bonus is a variable in the second term ETS equation. Its rationale is that airmen induced by a bonus to stay for a second term will be, on average, more likely to leave at the end of the second term than airmen who chose to stay without the inducement of a bonus. A more subtle form of this same argument would be to claim that the larger the proportion of an entering cohort that has persisted in the service to a given decision point, the higher the loss rate for the group

will be at that decision point. This phenomenon arises in the Gotz-McCall (1984) model of loss behavior. No such effect was found in these data, but a longer data series might uncover it; it should be checked for in the updating of the loss equations.

The bonus variables in the YAR file indicate the bonus an airman is eligible for in the airman's own AFSC. However, airmen can also receive bonuses for reenlisting if they retrain into an AFSC that is bonus-eligible. The cross-bonus variable that appears in the current first term ETS loss equation captures the effect of such retraining opportunities on losses. Since retraining from some specialties is unlikely, the cross-bonus variable weighs bonuses in other AFSCs by the historical probability that airmen do change from their current AFSC to the AFSC with a bonus. Updating the cross-bonus variable requires updating both the transition probabilities and the bonus opportunities. One task for the updating exercise will be to see if cross-bonus opportunities should be included in the second term ETS loss equations.

**Air Force Policy Periods.** Losses, extensions, reenlistments, and retirements are all sensitive to Air Force policies. The current first term ETS equations include variables for the period of Regular Reenlistment Bonuses and the period of operational manning policies. Other policy periods, such as the period during which extensions for personal reasons were permitted, could be incorporated into the model. The great tension is that the limited number of years of data can be easily overfit if too many main effects for such policy changes are permitted.

Careful treatment of policy periods is essential to maintaining the interpretation of the estimated loss equations as reflecting the desires of airmen to stay or leave (i.e., the supply of airmen). Losses are also influenced by the Air Force's demand for airmen; if this demand is not controlled for by the independent variables in the model, the estimated relationships become a confused mix of supply and demand. For example, the introduction of career gates in the first term is a change in the Air Force's demand for airmen that will raise loss rates even if airmen's willingness to reenlist is unchanged. Such a policy change must be controlled for if the supply of airmen is to be estimated from the available data.

Perhaps more important than incorporating policy periods into the model when estimating the equations' parameters is remaining aware of current policy changes that might cause the historical loss equations to systematically misforecast loss rates. Users of the EFMS must be alerted to the inability of the loss equations to anticipate the effects

of many policy changes. The models should be able to predict the effects of altered bonuses, pay, and employment opportunities, but many other policy changes must be accounted for by knowledgeable adjustments of the loss equations' predictions.

One loss equation's current form most intensely reflects this caution. The first term extend-given-stay equation currently has a specific effect for each fiscal year in the data and has no pay or unemployment variables in it. The only economic effect explicitly in the model is a bonus variable. The model is impoverished in specification because Air Force policies over the years have dominated first term extension behavior. The period of operational manning and the period of personal purpose extensions are too confounded with pay and unemployment in our sample to permit plausible estimation of those economic variables' influences. However, the data are rich enough to permit reliable estimates of the bonus effects, so the development of a statistical extend-given-stay model is warranted.

To implement the first term extend-given-stay model, the user must choose a base level for the extension rate. In essence, the user must decide which past fiscal year is the best model for the forecast years. Obviously, such judgments require good knowledge of past and anticipated extension policies. The first term extend-given-stay equation is unique in that it cannot be implemented without such judgments. *But in fact, none of the equations should be used without first considering whether similar judgments are needed to modify the equations' forecasts to reflect anticipated changes in Air Force policies.*

Once a base level for the extension rate is established, the extend-given-stay model can be used to assess the effects of alternative bonus policies on extensions. This important exercise need not rely on statistically estimated coefficients for pay and unemployment.

### **Estimating the Models**

Two considerations are paramount in choosing estimation techniques for the middle-term loss equations: (1) The coefficients on economic variables, particularly bonus variables, should be estimated without bias, and (2) the AFSC-specific effects estimated should avoid spurious variation arising from small cell sizes.

The first consideration dictates that the estimated model be richly specified so that biases from omitted variables are made unlikely. The second consideration calls for parsimony in the specification of AFSC effects.

The tension between rich specification and parsimony is resolved by estimating and implementing the loss equations in stages. In the first stage, parsimony is ignored. All likely correlates of loss rates are included in the equations, even those that do not appear in the IPM structures. In the first and second term ETS models, this principle is taken to the extreme of permitting each AFSC to have its own specific main effect in the loss equations. The estimates of parameters for bonuses and other variables are thereby subjected to minimum risk of bias.

In the second stage, the estimated effects for AFSCs are reconsidered. Implicitly it is assumed that airmen in the same two-digit AFSC category have similar loss rates, so that low variance estimators of a two-digit category have smaller mean square error as estimators of specific AFSC effects than do high variance estimators based only on data from the AFSC (without shred) itself. The initial least squares estimators of specific AFSC effects are therefore used only when the number of observations in the AFSC is large (say 50 or 75 or 100). For smaller AFSCs, the mean of the initially estimated AFSC effects for the AFSC's two-digit category is used as the estimate of the AFSC's effect. If the two-digit category is too small, the overall mean AFSC effect is used for each AFSC in that two-digit category. This second stage improves the mean square error of predictions both across AFSCs and for individual AFSCs.

In the third stage, the equations are simplified for use in the IPMs. Variables that are not tracked in the IPMs are set to a constant value or to a value that is constant within each AFSC, and their resulting contribution to loss rates is added to the constant term or to the AFSC-specific effects. The constant value chosen for these variables may be a recently observed value (for the Air Force or for each AFSC) or a historical mean value, whichever the analyst believes is the better guess for what the value will be over the forecast period.

The first and third stages are designed to provide unbiased forecasts of the effects of changes in military compensation, including bonuses, while maintaining relative parsimony in the cell structures of the IPMs. The strategy is most reliable when demographic or occupational variables are "folded into" the constant term, because the demographic and occupational structure of the force is unlikely to change substantially over the forecast period of the middle-term IPMs. However, the strategy is also applied to the cross AFSC bonus effect (the variable WBONC, which is documented in Trautman, 1986), and the stability of this variable over forecast periods is problematic.

When the loss equations are updated, some checking of the efficacy of assuming a fixed WBONC by AFSC is warranted.

The staged approach to estimating and implementing the loss models does not address three potential statistical problems: (1) multicollinearity among the economic variables, (2) serial correlation in loss and extension rates across the sample years, and (3) endogeneity of some explanatory variables such as grade. The data used to estimate these models originally (the estimates reported in Carter et al., 1987) suffered the first of these problems, could not support an analysis of the second, and did not evidence the third. With additional years of data, the first problem is likely to fade and the others to become more readily analyzed. Assessing the performance of updated specifications will require attention to each of these potential problems, but more generally, diagnostic examination of the equations' fits over time will assist in settling upon an appropriate specification. Appendix A illustrates the updating process. It describes how the first term ETS loss prediction model was updated from the specification in Carter et al., 1987.

**Multicollinearity.** Multicollinearity among the economic variables arises from the small number of years in our sample coupled with the nature of the economic variables themselves. The sample for the current estimates of the middle-term equations reported in Carter et al., 1987, is only ten years long, July 1973–June 1983 (and the first term equations can rely on only the last seven years, since the first three years' data are strongly influenced by the draft). These ten years of data provide the only variation in economic variables with which parameters can be estimated. There is no variation across individuals in either the military/civilian pay ratio or the unemployment rates. And, while bonuses vary from one AFSC to another, the AFSC-specific constants in the equations cause only within-AFSC variation in bonuses—i.e., temporal variation—to influence the estimated bonus effect.

Multicollinearity leads to imprecise parameter estimates for the collinear variables. But the more serious problem that can arise from multicollinearity is that the model may be seriously overfitted and spurious relationships introduced into the specification.

The multicollinearity problem is exacerbated by the variables that indicate specific policy periods across the sample years. For example, an initial specification of the first term extend-given-stay equation included variables for the period in which Regular Reenlistment Bonuses were given and for the period in which extensions for personal

reasons were permitted. (The latter was interacted in the model with the pay and unemployment variables.) The test and evaluation exercise revealed that the model was severely overfitted<sup>2</sup> and that, as a consequence, the unemployment coefficient was almost certainly badly estimated. In the current, revised specification of this model, overfitting is avoided, but the collinearity problem prevents estimation of separate pay and unemployment effects.

More years of data will alleviate the multicollinearity problem, but the risk of overfitting the loss equations when updating will persist for some time. *Identifying overfitted equations is a major purpose of the test and evaluation exercise.* The economic and policy period variables are most subject to overfitting; these should be given particular attention when evaluating the equations.

**Serial Correlation.** Serial correlation in loss rate disturbances of the models does not introduce biases in the estimated parameters, but it does reduce the efficiency of ordinary least squares as an estimation technique. The original equations were estimated without taking into account serial correlation. An attempt was made to estimate the statistical structure of the year-to-year correlations among the disturbances so that it would be possible to use generalized least squares estimation techniques. However, the limited number of years in the current sample precluded such joint estimation of correlations and equation coefficients; such estimations yielded implausible parameter estimates for numerous variables in the equations. This outcome is not surprising because generalized least squares estimators are theoretically inferior to OLS when the correlations cannot be estimated precisely.

In updating the loss equations, further attempts to incorporate serial correlation in the disturbances should be made. A three-step procedure for doing this is: (1) use individual data to estimate the parameters of a loss equation, (2) use fitted observations grouped by year to estimate the serial correlations among years, and (3) with an estimate of serial correlation in hand, return to the individual data to reestimate the equation parameters with generalized least squares (conditioned on the serial correlation estimates). Iterations of such a procedure will yield maximum likelihood estimates of the equation's parameters. Whether such a procedure yields better forecasts than OLS

---

<sup>2</sup>The numerous policy period dummies had invited extensive data mining for a best fitting set of interactions among the dummies and other variables in the model. The validation exercise that tested the original specification with out-of-sample data revealed that the specification incorporated apparently reflected spurious interactions.

depends on how well the serial correlations are estimated. With more years of data, generalized least squares will eventually be preferable to OLS. (Generalized least squares is always superior to OLS if correlations are known; the puzzle arises because the correlations must be estimated.)

The simplest form of serial correlation to correct for is a first order autoregressive disturbance. To test whether this is the appropriate form for the loss models, the autocorrelation function of the mean annual OLS residuals should be analyzed. Any textbook of time series analysis will contain a detailed discussion of the autocorrelation function. (Appendix A reports such an analysis from the first updating exercise conducted as part of the EFMS.)

**Endogeneity.** An explanatory variable whose value arises independently of the dependent variable in an equation is called "exogenous." An explanatory variable whose value is influenced by the dependent variable is called "endogenous." Ordinary least squares yields unbiased estimators of an equation's parameters only if all the explanatory variables are exogenous.

The endogeneity of an airman's circumstances and of Air Force policy must be considered when estimating the equations of these models because such endogeneity may result in inappropriate parameter estimates. For example, higher estimated loss rates for lower grades may not imply that higher promotion rates would lead to lower average loss rates. Airmen planning to stay in the service may study harder for promotion tests and therefore achieve higher scores than their colleagues planning to leave the service. Similarly, high loss rates in AFSCs receiving bonuses (which are often observed) does not imply that raising bonuses will raise losses. The Air Force often gives bonuses to AFSCs precisely because their loss rates are especially high.

The current equation estimates account for the potential endogeneity of bonuses by not using across AFSC variation in loss rates to estimate the effect of bonuses. AFSC-specific constant terms accomplish this. However, the bonus effect coefficient estimates are not entirely purged of potential endogeneity. If Air Force bonus allocations vary from year to year in response to expected fluctuations in loss rates (say in response to civilian economic opportunities), then the year to year variations in bonuses even within AFSCs will introduce some bias into the estimates of the effects of bonuses on losses.

Instrumental Variables is an estimation technique that could in principle avoid all the biases from endogeneity, but analysts are quite unlikely to find suitable instruments. Variables correlated with the level of bonus allocations are likely also to be correlated with the determinants of airmen's civilian opportunities. Rather, the bias reduction obtained by including civilian opportunities explicitly in the equations will probably have to be sufficient.

The current estimates do not reflect corrections for the endogeneity of grade. Since promotion rates are an important policy tool available to the Air Force, safeguards against biases in grade effects are needed. Instrumental Variables (IV) estimators for the grade parameters in the second term loss equations were computed. The hypothesis of no endogenous grade effect could not be rejected, so the OLS estimators were used because they are more efficient if there is no bias. Because the only instrument for grade is the proportion of one's entering cohort who are promoted, yet another variable that only changes over time, the addition of more years of data may expose an endogenous grade effect. The techniques suggested by Carter et al., 1987, for testing for an endogenous grade effect should be repeated in the updating exercise. If a bias is found, IV estimators for all the coefficients in the equations should be used. (The test suggested by Carter et al., 1987, obtains IV estimates for the grade parameters but may obtain biased estimates for the remaining parameters. It is suitable for testing for bias but may not be suitable for correcting for bias.)

**Diagnostics.** The signs and magnitudes of estimated coefficients and their standard errors are one guide to the appropriateness of a specification. But it is also important that the fit of the updated model across groups and especially over time be examined carefully. The computer programs that compute the AFSC-specific constant terms also produce summary statistics, such as the mean residual for each fiscal year and the mean residual for each AFSC in each fiscal year. If the equations do not fit all years comparably well, one must ask if there is an overlooked misspecification. If the performance of the model deteriorates at one end of the sample or the other, one must wonder if there has been a change in regime that requires dropping observations or adding policy period variables. Careful attention to sample sizes from one fiscal year to the next can also uncover bugs in the sample selection code that could otherwise lead to unnoticed biases in the estimated parameters. (Abrahamse, 1988, provides an exemplary use of diagnostics to assess the middle-term loss models.)

### **Policy Uses and the Middle-Term Loss Equations**

The middle-term loss equations are shaped by the policy uses of the EFMS. The loss models must be able to forecast the effects of changes in the policies most important to the users of the EFMS. For some complex policy changes, like altering the retirement system, the accommodation of the middle-term loss equations will be complicated. For adaptation to other new policy questions, such as a concern about the effects of a change in the CONUS/non-CONUS mix of assignments for airmen, the changes in the model will be simpler. But in all cases, *the guiding principle is to make the loss equations as responsive as possible to the policy questions being addressed by the EFMS.*

Losing track of the policy requirements of the system is not difficult. For example, when the original first term extend-given-stay equation was found in testing and evaluation to have been badly overfitted, it was necessary to respecify the equation. The first revised specification that was suggested dropped all economic variables from the model, even though important bonus effects in the extend-given-stay decision had been identified in the initial specification. The focus on overcoming the statistical problems found in the initial specification distracted attention, although not for long, from the specification of a policy-relevant equation. As the loss equations are updated, constant vigilance will be the only way to maintain both statistical legitimacy and policy relevance in the equations' specifications.

Maintaining this balance will sometimes be easy. For example, as more years of good CONUS/non-CONUS data become available, the addition of this variable to the loss equations will allow the system to answer questions about CONUS balance. But at other times, the balance will be hard to maintain. For example, as the new retirement system begins to influence airmen's decisions, the estimated loss models will incompletely capture the change in regime. Only after considerable experience with the new regime will updated specifications of the loss models reflect the new conditions accurately. In the meantime, ad hoc adjustments to the forecasts of the loss equations, adjustments based on careful use of Arguden's (1986) retirement model, will be needed to avoid forecast biases.

### **A Summary of Future Prospects**

Updating the middle-term loss equations offers prospects of numerous improvements in the forecasting accuracy of the equations:

1. Less multicollinearity among the economic variables and therefore more precise estimates of the effects of pay and unemployment.
2. More precise estimates of AFSC-specific effects and therefore lower mean squared forecast errors in the disaggregate IPMs.
3. Inclusion of CONUS variables in the equations and therefore an extended policy capability of the EFMS.
4. A more reliable assessment of the exogenous effects of changes in promotion policies and therefore a further enhanced policy capability of the EFMS.
5. Incorporation of serial correlation in loss rates into the estimation procedures and an accompanying increase in forecast efficiency.
6. A more reliable assessment of cross-bonus effects in both the first and second terms, adding to the quality of bonus allocation analyses in the EFMS.

The equation that stands to improve most is the first term extend-given-stay equation. Currently it relies on a user's judgment about extension policies in place to set a base level of extensions. With more years of data, the need for informed guessing may be lessened.

The one storm cloud on the horizon is the gathering effect of the new retirement system on airmen's decisions. The full force of those changes will not strike for ten years or so; but when it does, there will be a period of transition during which the second term, career, and retirement models may perform poorly if not adjusted with corrections based on Arguden's model. The first term equations are likely to be little affected by the new retirement policy.

## **Appendix A**

### **UPDATING THE FIRST TERM ETS MIDDLE-TERM LOSS PREDICTION MODEL<sup>1</sup>**

#### **INTRODUCTION**

The Enlisted Force Management System uses a family of middle-term loss prediction models for forecasting losses one to six years into the future. The original specifications of those models are documented in Carter et al., 1987. Those specifications used data through June of 1983. An integral part of the EFMS will be the periodic updating of these (and other) models as additional data become available. This appendix illustrates that updating process; it documents the updating of the middle-term loss model for first term airmen who are making the decision to stay in the service or leave the service during the year just preceding their originally scheduled expiration of term of service. The update uses data through May of 1985.

The first term ETS model described in Carter et al., 1987, Sec. V, forecast poorly outside the sample period. The predicted losses for FY 85 exceeded actual losses by 20 percent. Air Force analysts speculated that the problem lay in the model's unemployment coefficient. The model forecast large changes in loss rates in response to the large changes in unemployment during the forecast period; these changes simply did not materialize. The analysts speculated that a change in airman behavior had occurred that might be best captured in an updated model by weighing more recent observations more heavily than earlier observations.

The analysts' first speculation is correct, but their proposed solution is incorrect. New exploratory analyses, spurred by the findings of the Air Force analysts, indicate that the original ETS loss model was misspecified. Correcting this misspecification does result in a smaller coefficient for unemployment and will, one hopes, rectify the forecasting problems of the model.

The coefficients for economic variables in the original specification were clearly at risk of having been misestimated. There were only seven years of data available for fitting the first term ETS model, from July of 1976 through June of 1983. These few

---

<sup>1</sup>Erik Jamryd and Alan Siqueira, students at Bates College, provided able and conscientious assistance.

Table A.1

COMPARISON OF ORIGINAL SPECIFICATION OF FIRST TERM ETS  
LOSS MODEL USING DATA FROM YAR 2.75 AND YAR 3.0<sup>a</sup>  
(Using data for the period July 1976-June 1983)

Predictor Variable	Variable Name	YAR 2.75		YAR 3.0	
		Coeff.	t-stat.	Coeff.	t-stat.
Male	BMALE	0.137	(19.91)	0.138	(19.49)
Married	MAR	-0.025	(-2.85)	-0.033	(-3.76)
Male and married	MAR*BMALE	-0.089	(-9.50)	-0.077	(-8.07)
Black	BBLK	-0.172	(-37.48)	-0.173	(-37.17)
Log(mil wages/civ wages)	LMCPAY	-0.437	(-4.68)	-0.731	(-7.11)
4-year $\times$ log(moving average of unemployment)	T4*LOGMAU	-0.361	(-28.94)	-0.291	(-20.08)
Six-year enlistee	T6	0.685	(24.62)	0.530	(16.35)
Half bonus	BH	0.001	(0.06)	-0.047	(-2.28)
Bonus level=1	B1	-0.034	(-5.60)	-0.034	(-5.20)
Bonus level>1	BG1	-0.013	(-1.89)	-0.001	(-0.24)
Cross-bonus average	WBONC	-0.022	(-1.55)	-0.027	(-0.88)
Period of Regular Reenlistment Bonus	BRRB	0.078	(15.79)	0.067	(13.86)
Period of operational manning	OPMAN	-0.030	(-6.75)	-0.030	(-6.86)
Mean rate		.480		.482	
Sample size		95069		89453	

<sup>a</sup>Records in which the cross-bonus variable (WBONC) is missing have been deleted.

The demographic coefficients do not differ importantly; only married females have a change greater than .004, and for them the change is still less than .01. However, the economic variables' coefficients do differ appreciably between the two samples. The differences are twice troubling. First, replicating the previous specification with a new YAR but the old sample period should be a valuable consistency check for unearthing problems in a new YAR that might otherwise go undetected. The value of this consistency check is great enough to demand that either base years for ultimate AFSCs be kept fixed from one YAR to the next or models always be estimated and documented with and without AFSC controls, so that the latter estimates would be available for consistency checks when new YARs are available.

The second troubling implication of the different estimates is more substantive. The reason for controlling for ultimate AFSC in the first place is to keep occupational differences from spuriously influencing the coefficients for other variables, especially the bonus variables. If shifting the base year for the ultimate AFSCs substantially changes the economic variables' coefficients, ultimate AFSCs in one year or the other—or more likely both—are not truly controlling for occupation. This raises the specter that the bonus coefficients are biased, a potentially serious problem. Poorly specified occupation categories can appreciably alter bonus coefficient estimates, but the problem does not appear to be serious in the updated model. Nonetheless, an important agenda item for the development of the EFMS is the establishment of accurate and temporally consistent occupation designators for use in future updates of the middle-term loss models. (See the discussion of Airmen's Economic Opportunities in Sec. III.)

Analysts who used the original specification of the first term ETS loss models noted that while early releases were counted among the separations at ETS, early reenlistments were not. The omission of early reenlistments was inadvertent, and their inclusion in the updated model is desirable. Table A.2 shows the parameter estimates for the original specification and the original sample period (using YAR 3.0) for samples without and with early reenlistments. The most notable changes in the estimates are in the variables for bonuses greater than one and for cross-bonus opportunities. The early reenlistments are disproportionately among airmen with high bonuses, and airmen cross training into new occupations that have bonuses. The only other variable affected much is the military/civilian wage ratio. The early reenlisters are included in all subsequent samples in the updating exercise.

The ultimate AFSC variable poses yet another problem for estimating the first term ETS loss model. Airmen's bonus opportunities in AFSCs other than their own are measured by the structure of bonuses across AFSCs and by the historical probabilities of an airman moving from one AFSC in his first term to another in his second term. Transitions between ultimate AFSCs are not always well defined, so some airmen do not have a value for the cross-bonus opportunity variable (WBONC). In YAR 2.75, some 5000 airmen were not included in the estimation because they did not have values for WBONC. An alternative approach would be to set WBONC equal to zero when it is missing and to add a dummy variable (WBONCM) indicating cases for which this had been done. Table A.3 indicates that including coefficients; only the military/civilian

Table A.2

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
EXCLUDING AND INCLUDING EARLY REENLISTMENTS<sup>a</sup>  
(Using YAR 3.0 data for the period July 1976-June 1983)

Predictor Variable	No Early Reups		With Early Reups	
	Coeff.	t-stat.	Coeff.	t-stat.
Male	0.138	(19.49)	0.139	(19.82)
Married	-0.033	(-3.76)	-0.032	(-3.67)
Male and married	-0.077	(-8.07)	-0.079	(-8.35)
Black	-0.173	(-37.17)	-0.172	(-37.15)
Log(mil wages/civ wages)	-0.731	(-7.11)	-0.879	(-8.64)
4-year $\times$ log(moving average of unemployment)	-0.291	(-20.08)	-0.287	(-19.91)
Six-year enlistee	0.530	(16.35)	0.511	(15.85)
Half bonus	-0.047	(-2.28)	-0.045	(-2.17)
Bonus level=1	-0.034	(-5.20)	-0.033	(-5.10)
Bonus level>1	-0.001	(-0.24)	-0.019	(-3.04)
Cross-bonus average	-0.027	(-0.88)	-0.044	(-1.47)
Period of Regular Reenlistment Bonus	0.067	13.86	0.061	(12.76)
Period of operational manning	-0.030	(-6.86)	-0.028	(-6.36)
Mean rate	.482		.475	
Sample size	89453		90712	

<sup>a</sup>Records in which the cross-bonus variable (WBONC) is missing have been deleted.

wage ratio variable shows much effect at all. (The missing WBONC cases are limited to a few ultimate AFSCs, and all airmen in those ultimate AFSCs have WBONC missing. Consequently, WBONCM is collinear with the AFSC dummies.) The cases with WBONC missing are included in subsequent analyses.

Another oversight in the original sample for estimating the first term ETS loss model was the inclusion of six-year enlistees who entered the service between July 1970 and June 1972. Four-year enlistees who entered the service during this period were excluded from the sample because the influence of the draft on their choices was thought to be too strong. This consideration led to the choice of July 1976 as the first year of original ETS to be included in the sample. To be consistent, the first year of original ETS

Table A.3

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
EXCLUDING AND INCLUDING CASES MISSING WBONC  
(Using YAR 3.0 data for the period July 1976-June 1983)

Predictor Variable	Missing WBONC Out		Missing WBONC In	
	Coeff.	t-stat.	Coeff.	t-stat.
Male	0.139	(19.82)	0.135	(20.05)
Married	-0.032	(-3.67)	-0.034	(-4.05)
Male and married	-0.079	(-8.35)	-0.080	(-8.71)
Black	-0.172	(-37.15)	-0.170	(-37.94)
Log(mil wages/civ wages)	-0.879	(-8.64)	-0.934	(-9.79)
4-year $\times$ log(moving average of unemployment)	-0.287	(-19.91)	-0.289	(-20.97)
Six-year enlistee	0.511	(15.85)	0.516	(16.74)
Half bonus	-0.045	(-2.17)	-0.043	(-2.40)
Bonus level=1	-0.033	(-5.10)	-0.032	(-5.54)
Bonus level>1	-0.019	(-3.04)	-0.019	(-3.30)
Cross-bonus average	-0.044	(-1.47)	-0.043	(-1.46)
Cross-bonus average missing Period of Regular Reenlistment	(cases deleted)		0.0	collinear
Bonus	0.061	(12.76)	0.062	(13.23)
Period of operational manning	-0.028	(-6.36)	-0.026	(-6.28)
Mean rate	.482		.473	
Sample size	89453		100479	

to be included for six-year enlistees should have been July 1978. Table A.4 shows that this choice has little effect on the original specification. Six-year enlistees entering the service before July 1972 are deleted from subsequent estimations.

The preceding four adjustments to the original analysis (using the YAR 3.0 file, including early reenlistment cases, including cases missing WBONC, and deleting six-year enlistees who entered the force before July 1972) led to a reestimation of the original model with nearly two additional years of data (July 1984 through May 1985). Table A.5 presents the estimates using the smaller and larger samples.

The enlarged sample leads to no appreciable changes in the demographic coefficients. The military/civilian wage ratio coefficient is markedly reduced, however,

Table A.4

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
EXCLUDING AND INCLUDING SIX-YEAR ENLISTEES  
WHO ENLISTED BEFORE JULY 1974  
(Using YAR 3.0 data for the period July 1976-June 1983)

Predictor Variable	Include Early TOE=6		Drop Early TOE=6	
	Coeff.	t-stat.	Coeff.	t-stat.
Male	0.135	20.05	0.135	(20.05)
Married	-0.034	-4.05	-0.034	(-4.05)
Male and married	-0.080	-8.71	-0.079	(-8.70)
Black	-0.170	-37.94	-0.170	(-37.94)
Log(mil wages/civ wages)	-0.934	-9.79	-0.932	(-9.76)
4-year $\times$ log(moving average of unemployment)	-0.289	-20.97	-0.290	(-20.99)
Six-year enlistee	0.516	16.74	0.518	(16.77)
Half bonus	-0.043	-2.40	-0.043	(-2.40)
Bonus level=1	-0.032	-5.54	-0.032	(-5.55)
Bonus level>1	-0.019	-3.30	-0.019	(-3.27)
Cross-bonus average	-0.043	-1.46	-0.035	(-1.15)
Cross-bonus average missing	0.0	collinear	0.0	collinear
Period of Regular Reenlistment				
Bonus	0.062	13.23	0.062	(13.26)
Period of operational manning	-0.026	-6.28	-0.026	(-6.34)
Mean rate		.473		.473
Sample size		100479		100302

and the "half bonus" and "bonus equal to one or more" coefficients are markedly increased. The unemployment coefficient behaves perversely, from the perspective of those analysts who thought the coefficient too high in the original specification and responsible for the too volatile forecasts over the years now added to the sample: The estimate is even higher in the expanded sample. But the most telling change is in the two temporal dummy coefficients. If the model were properly specified, much of the drop in mean loss rates from the original sample (.47 to .45) would probably be captured by the explanatory variables of the model. Instead, the two dummies shift about 2 percentage points apiece, suggesting that the model may be missing some temporal evolution of loss behavior.

Table A.5

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
FOR ORIGINAL SAMPLE PERIOD (7607-8306) AND  
ENLARGED SAMPLE PERIOD (7607-8505)  
(Both specifications use YAR 3.0 Data)

Predictor Variable	Variable Name	Old Sample		New Sample	
		Coeff.	t-stat.	Coeff.	t-stat.
Male	BMALE	0.135	(20.05)	0.120	(20.71)
Married	MAR	-0.034	(-4.05)	-0.029	(-4.03)
Male and married	MAR*BMALE	-0.079	(-8.70)	-0.087	(-10.97)
Black	BBLK	-0.170	(-37.94)	-0.163	(-41.33)
Log(mil wages/civ wages)	LMCPAY	-0.932	(-9.76)	-0.530	(-6.16)
4-year $\times$ log(moving average of unemployment)	T4*LOGMAU	-0.290	(-20.99)	-0.337	(-31.16)
Six-year enlistee	T6	0.518	(16.77)	0.617	(25.08)
Half bonus	BH	-0.043	(-2.40)	-0.058	(-6.60)
Bonus level=1	B1	-0.032	(-5.55)	-0.041	(-8.04)
Bonus level>1	BG1	-0.019	(-3.27)	-0.014	(-2.76)
Cross-bonus average	WBONC	-0.035	(-1.15)	0.033	(1.18)
Cross-bonus average missing	WBONCM	0.0	collinear	0.0	collinear
Period of Regular Reenlistment					
Bonus	BRRB	0.062	(13.26)	0.095	(23.52)
Period of operational manning	OPMAN	-0.026	(-6.34)	0.001	(0.51)
Mean rate		.472		.451	
Sample size		100302		127157	

The documentation of the original specification presented evidence of some temporal instability in the estimates of the economic variables. Carter et al., 1987, Table 5.3, shows that dropping earlier years from the sample leads to larger estimated effects for both pay and unemployment. Table A.6 reproduces that table. It presents the equations of the loss model based on three alternative sample periods and for the full sample. Demographic effects were very stable across the samples, and bonus effects showed modest mixed changes. Carter et al. note that the parameter instability was to be expected given the very small number of time periods over which the model was being estimated.

The most striking feature of Tables A.5 and A.6 is that they conflict with the expectations of analysts who believe that, if anything, estimated economic coefficients should be smaller in the expanded sample and in samples deleting earlier observations.

Table A.6

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS  
MODEL WITH FITS FOR FULL SAMPLE PERIOD  
AND FOR THREE SUBPERIODS  
(Using YAR 2.75 data)

Predictor Variable	7807- 8306 Coeff.	7907- 8306 Coeff.	8010- 8306 Coeff.	7607- 8306 Coeff.
Male	0.115	0.112	0.103	0.137
Married	-0.036	-0.033	-0.032	-0.025
Male and married	-0.073	-0.079	-0.095	-0.089
Black	-0.167	-0.161	-0.157	-0.172
Log(mil wages/civ wages)	-0.415	-0.717	-0.908	-0.437
Log(moving average of unemployment) if four-year enlistee	-0.392	-0.368	-0.483	-0.361
Six-year enlistee	0.764	0.716	0.950	0.685
Half bonus	0.005	-0.002	0.017	0.001
Bonus level=1	-0.034	-0.037	-0.028	-0.034
Bonus level>1	-0.019	-0.027	-0.043	-0.013
Cross-bonus average	0.012	-0.015	-0.000	-0.022
Period of Regular Reenlistment Bonus	0.068	0.044	0.000	0.078
Period of operational manning	-0.027	-0.014	-0.030	-0.030
Sample size	(70881)	(56828)	(36845)	(95069)

SOURCE: Carter et al., 1987.

However, the available data do not reject the contention of smaller economic effects out of hand. One specification tried in the original modeling exercise led to much smaller economic coefficients, especially for the unemployment variable. Inclusion of a time trend in the model lowered all the estimated economic coefficients in magnitude. (The demographic effects were hardly affected.) Table A.7 presents the original specification with and without time trend estimated over the original sample period. (Cases missing WBONC are deleted from these estimations. This does not alter the comparison noticeably.)

The time trend specification was dismissed from consideration in the original analysis for three reasons. First, the small number of years in the original sample already

Table A.7

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
WITH AND WITHOUT A TIME TREND<sup>a</sup>

(Using YAR 3.0 data for the period July 1976-June 1983)

Predictor Variable	Variable Name	No Time Trend		With Time Trend	
		Coeff.	t-stat.	Coeff.	t-stat.
Male	BMALE	0.138	(19.49)	0.133	(19.95)
Married	MAR				
Male and married	MAR*BMALE	-0.077	(-8.07)	-0.078	(-8.16)
Black	BBLK	-0.173	(-37.17)	-0.172	(-36.99)
Log(mil wages/civ wages)	LMCPAY	-0.731	(-7.11)	-0.589	(-5.71)
4-year $\times$ log(moving average of unemployment)	T4*LOGMAU	-0.291	(-20.08)	-0.095	(-4.80)
Six-year enlistee	T6	0.530	(16.35)	0.147	(3.52)
Half bonus	BH	-0.047	(-2.28)	-0.023	(-1.11)
Bonus level=1	B1	-0.034	(-5.20)	-0.029	(-4.55)
Bonus level>1	BG1	-0.001	(-0.24)	0.000	(0.13)
Cross-bonus average	WBONC	-0.027	(-0.88)	-0.001	(-0.05)
Period of Regular Reenlistment Bonus	BRRB	0.067	(13.86)	-0.036	(-4.18)
Period of operational manning	OPMAN	-0.030	(-6.86)	-0.003	(-0.73)
Time Trend, 7606=1	TREND83	—	—	-0.002	(-14.55)

<sup>a</sup>Records in which the cross-bonus variable (WBONC) is missing have been deleted.

ran risks of overfitting the model and adding a time trend could exacerbate this problem. Second, a time trend in a forecasting model, especially a model forecasting probabilities, poses serious problems. Does one expect the trend to continue indefinitely? For only a short time? To implement forecasts with a time trend of 2.4 percent per year (as estimated in the time trend specification) was likely to yield bad forecasts somewhere over the six-year horizon envisioned for this model.

But these objections to the time trend specification must be reconsidered in the light of the poor performance of the original specification, seemingly because of the high estimated coefficient on unemployment. The objections should be reconsidered, but not ignored. They compel us to examine the data for corroboration of the contention that the lower economic effects estimated in a model with a time trend are indeed the more appropriate estimates. Without such corroboration, the objections to including the time

trend should not be overruled. That examination—and the corroborating evidence it provides—are presented below.

One last specification issue is that changes in the ultimate AFSC categorizations can alter coefficient estimates for bonus variables. Table A.8 presents two versions of the original specification estimated for the expanded sample period. In one model, ultimate AFSC is controlled for (by adding a dummy variable for each ultimate AFSC). In the other, ultimate AFSC is altered slightly and the altered occupational code is controlled for when the model is estimated. The alteration is to split airmen in AFSC 511X0 off from others with whom they had been lumped in a single ultimate AFSC. These airmen are computer specialists who were, for arcane reasons, lumped together with a group of clerks. Shifting the ultimate AFSC for these fewer than 1000 airmen

Table A.8

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL CONTROLLING FOR  
ULTIMATE AFSC AND THE MODIFIED OCCUPATIONAL CODE  
(Using YAR 3.0 data for the period July 1976–May 1985)

Predictor Variable	Variable Name	Control For ULTAFC		Split 511X0 Off	
		Coeff.	t-stat.	Coeff.	t-stat.
Male	BMALE	0.120	(20.71)	0.120	(20.66)
Married	MAR	-0.029	(-4.03)	-0.029	(-4.05)
Male and married	MAR*BMALE	-0.087	(-10.97)	-0.086	(-10.96)
Black	BBLK	-0.163	(-41.33)	-0.163	(-41.19)
Log(mil wages/civ wages)	LMCPAY	-0.530	(-6.16)	-0.513	(-5.96)
4-year $\times$ log(moving average of unemployment)	T4*LOGMAU	-0.337	(-31.16)	-0.337	(-31.11)
Six-year enlistee	T6	0.617	(25.08)	0.616	(25.05)
Half bonus	BH	-0.058	(-6.60)	-0.060	(-6.80)
Bonus level=1	B1	-0.041	(-8.04)	-0.043	(-8.38)
Bonus level>1	BG1	-0.014	(-2.76)	-0.019	(-3.55)
Cross-bonus average	WBONC	0.033	(1.18)	0.034	(1.23)
Cross-bonus average missing	WBONCM	0.0	collinear	0.0	collinear
Period of Regular Reenlistment Bonus	BRRB	0.095	(23.52)	0.095	(23.46)
Period of operational manning	OPMAN	0.001	(0.51)	0.001	(0.49)
Mean rate	.451				
Sample size	12157				

results in a one-third increase in the magnitude of the coefficient on bonus levels higher than one. This sensitivity reinforces the contention that new, better occupational categories are needed for future updates of the loss models. (Fortunately, as noted below, the updated specification estimated over the expanded sample seems to be robust to changes in the occupational codes. The problem is more a threat to future specifications than to present specifications.)

### **A TEMPORAL ANALYSIS OF FIRST TERM ETS LOSSES**

In the original exploration of temporal effects in the first term ETS loss model, the temporal unit of analysis was a year (beginning in July of one calendar year and ending in June of the following calendar year). In the updating exercise, annual observations give way to monthly observations. The choice of monthly observations exploits the month to month variations in loss rates, unemployment, and bonuses when estimating coefficients, thereby adding many degrees of freedom to the analysis.

The temporal analysis has two objectives. First, it is desirable to determine if serial correlation among the disturbances of the model make OLS an inefficient estimation technique for the ETS loss model. *Second, it is important to know whether the constant term or other parameters in the model vary over time.* These two issues are analyzed together because tests for serial correlation will be biased if the temporal structure of the parameters is misspecified. The analysis in this section uses data for four-year enlistees. Similar results were obtained for six-year enlistees, although their loss rate disturbances displayed less autocorrelation.

If these data were monthly loss rates and monthly averages for the independent variables in the loss models, it would be possible to examine the residuals from an OLS estimation of the model for signs of serial correlation. (The Durbin-Watson statistic is the most common check for serial correlation. Below a more general test is based on the autocorrelation function of the least squares residuals.) Since these data are individual airmen's records, there is a more complex problem. The most straightforward specification of possible serial correlation in these models is to assume that each airman's probability of loss has two components: One common to all airmen in a given month, and one unique to each airman in each month. The analysis here allows that these month-specific disturbances may be serially correlated.

One way to test for serial correlation in these data would be to form monthly averages of all the variables and fit the model with OLS and those averages. However, to

do this would be to waste much of the information contained in the data set. For some variables most of the information would be lost.

Demographic characteristics vary markedly among airmen who come to ETS. But across months, there is little variation in the demographic composition of these airmen. Consequently, if monthly averages were used to estimate the model most of the information about demographic differences in loss rates would be lost. To use monthly data would also sacrifice much of the information about bonuses as much of their variation is across individuals rather than across months. In contrast, to use monthly data would lose none of the information about unemployment rates or wages because these variables do not vary across airmen in a given month.

An alternative estimation strategy permitted retention of the cross sectional information in the data set and still could provide a computationally simple way to look for serial correlation.

There are two steps in this alternative strategy. The first step estimates the loss model by OLS using all of the data, but including in the model a dummy variable for each month.<sup>2</sup> The estimated coefficients for the monthly dummies reflect the cross sectional sample data on demographics and bonuses. The dummies are perfectly correlated with the unemployment, military/civilian wage ratio, and policy period variables, so the coefficients of these latter variables cannot be estimated in this first stage. The second step uses the estimated coefficients of the monthly dummy variables as dependent variables in a model whose residuals will permit testing for serial correlation in the ETS loss model.

If one believes the original ETS specification is correct, then the model that explains the estimated coefficients of the monthly dummies would include only the month-specific unemployment rate, the month-specific military/civilian wage ratio, and a monthly indicator for each of the policy periods specified in the ETS loss model. However, if one believes that the original specification is incorrect, one might argue that other variables belong in the model. Most important, one might argue for the inclusion of a time trend in the model.

---

<sup>2</sup>The SAS procedure ABSORB allows one to do this in a computationally efficient way. The procedure uses deviations from means as data to obviate the need for including the many dummies in the  $X'X$  matrix. The user must then compute the coefficients for the dummies in a second step. The coefficient estimator for each dummy is the mean residual for the group for whom the dummy is one.

Tests for serial correlation are biased when the independent variables of the model are misspecified. In particular, incorrect omission of a time trend will invalidate such tests. This problem is illustrated in the results reported in Table A.9. The first order autoregression coefficient ( $\rho$ ) estimate if we rely on the original specification is .46, but if we include a time trend in the model, the estimate is .28. (The difference between these estimates is large relative to their standard errors.)

As noted above, it did not seem a good idea to include a time trend in the model when there were few data. In the now modestly larger data set it still seemed unwise. Additional confirmation from the data, beyond the statistical significance of the time trend itself, was sought that the time trend belonged in the model. That additional confirmation was found.

Table A.9

ORIGINAL SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
CORRECTING FOR FIRST-ORDER SERIAL CORRELATION  
(Using YAR 3.0 data for the period July 1976–May 1985)

Predictor Variable	Autoregressive Ordinary Least Squares			
	With Time Trend		Without Time Trend	
	Coeff.	t-stat.	Coeff.	t-stat.
Constant	2.165	(1.97)	1.7010	(1.29)
Male	-0.081	(-0.44)	-0.1660	(-0.86)
Married	-0.014	(-0.04)	-0.2571	(-0.74)
Male and married	0.189	(0.53)	0.5209	(1.44)
Black	-0.121	(-1.23)	-0.0599	(-0.53)
Log(mil wages/civ wages)	-0.514	(-2.19)	-0.4352	(-1.54)
4-year $\times$ log(moving average of unemployment)	-0.173	(-5.57)	-0.2149	(-5.54)
Half bonus	0.132	(1.01)	-0.3812	(-5.07)
Bonus level=1	-0.019	(-0.43)	-0.1012	(-1.96)
Bonus level>1	-0.072	(-0.90)	-0.1466	(-1.64)
Regular Reenlistment Bonus	0.441	(29.91)	0.4845	(32.85)
Operational manning	-0.003	(-0.36)	-0.0176	(-1.56)
Time trend	-0.002	(-4.68)	—	—
	$\rho$	= 0.281	$\rho$	= 0.462
	t-stat	= (2.81)	t-stat	= (5.13)

The confirmation came from the failure of the model without the time trend to yield estimates of the bonus coefficients consistent with those obtained from the cross sectional data. Recall that the estimated coefficients of the monthly dummies are the monthly loss rates less the (fitted) effects of the demographic and bonus variables. Since the monthly dummies' coefficients are purged of the demographic and bonus effects, the coefficients of demographics and bonus effects have an expected value of zero if those variables are included in the model for the monthly dummies' coefficients—or they do if the loss model itself was properly specified. Consequently, if the monthly averages of the demographic and bonus variables are included in the second stage model and found to be significant, the model is misspecified. (This test is equivalent to comparing the demographic and bonus effect estimates from the first stage with estimates from regressing monthly loss rates on average values of the variables in the ETS loss model.)

Table A.9 reports the results of regressing the estimated monthly dummies' coefficients from stage one against the monthly averages of all the variables in the original ETS specification. The regressions are based on the Cochran-Orcutt procedure that corrects for the first order autoregression in the disturbances. Note that when a time trend is included in the model, the demographic and bonus coefficients are all statistically insignificant, but when the time trend is omitted, the bonus coefficients become significant.

There are three reasons for including the time trend in the updated model: (1) the time trend is statistically significant; (2) including the time trend lowers the unemployment coefficient, which would have improved the predictive performance of the original model; and (3) exclusion of the time trend induces an inconsistency between the bonus effects observed in cross sectional and intertemporal slices of the data. In the initial analysis, the problems of coping with a time trend in a forecasting model and the fear of incorporating a spurious temporal effect prevented the inclusion of the time trend despite having reason (1) in hand. The poor initial forecasts of the model drew attention to (2) and encouraged further exploration. The corroboration of reason 3 left little choice but to include a time trend in the model.

If subsequent updates of the model indicate that the time trend persists or proves volatile, efforts should be made to identify the causes of the trend so that they might be explicitly modeled. These causes may not be so much economic as social or institutional—i.e., the esteem of the military in the public mind or changes in the rules

that govern the everyday lives of military personnel. In the meantime, to accurately estimate the effects of changing economic circumstances, a time trend is necessary in the model.

The regressions reported in Table A.9 are equivalent to regressions using monthly averages for the loss rate as the dependent variable; they do not capitalize on the informational advantage that initially motivated use of the monthly dummies' coefficients as dependent variables. To get that advantage would require omitting the demographic and bonus variables from the model. However, the results of that exercise are not reported because they are not much different from the results given in Table A.9.

Figures A.1 and A.2 show unemployment and the military/civilian wage ratio plotted against the residuals from the model of Table A.9 that includes the time trend. The plots evidence no nonlinearities in the effects of these economic variables.

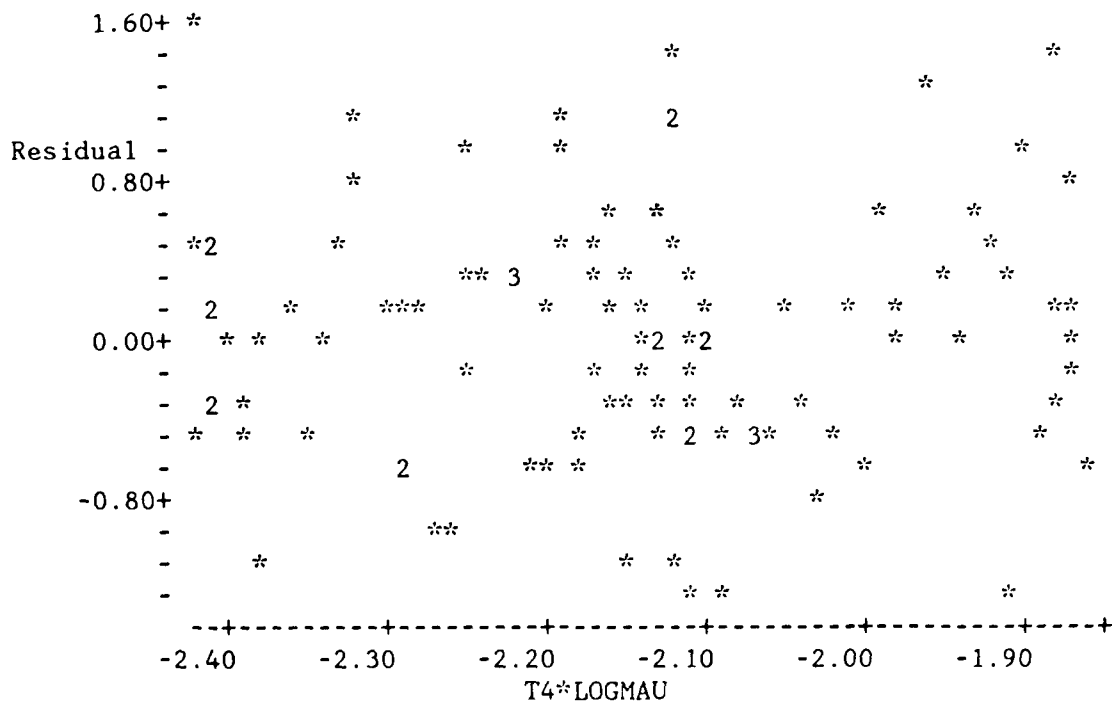
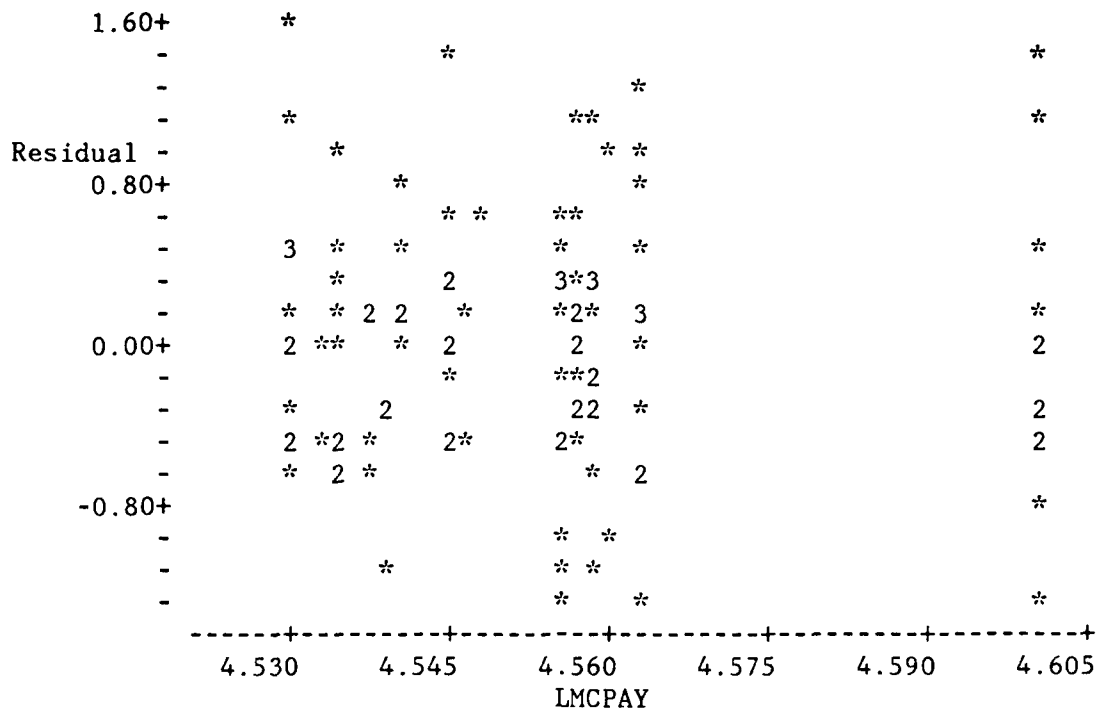


Fig. A.1—Plot of unemployment variable versus residuals from first order autoregressive time trend model of Table A.9.



**Fig. A.2—Plot of military/civilian wage ratio variable versus residuals from first-order autoregressive time trend model of Table A.9.**

Table A.10

**MODIFIED BOX-PIERCE  $\chi^2$  TEST STATISTICS FOR REMAINING  
SERIAL CORRELATION AT UP TO 12, 24, 36, AND 48 MONTHS  
AFTER CORRECTING FOR FIRST-ORDER SERIAL  
CORRELATION AMONG THE RESIDUALS OF  
THE AUTOREGRESSIVE MODELS  
(With and without time trend)  
(Using YAR 3.0 data for the period July 1976–May 1985)**

	Lag (Months)			
	12	24	36	48
With time trend	21.8	30.7	48.2	65.3
Without time trend	13.5	28.3	46.1	67.3
Degrees of freedom	12	24	36	48

Heartening to the users of the EFMS is the low degree of serial correlation found in the residuals ( $\rho=.28$ ). Since the autoregressive scheme implies an exponential dampening of serial correlation, the degree of serial correlation is below .1 after one month and below .01 after three months. Since the middle-term model is intended for forecasting losses beyond one year in the future, serial correlation need not be accounted for in making forecasts with the model. Table A.10 reports test statistics that confirm that no significant serial correlation appears in the residuals once first order autoregression is corrected for. (Validators of the model, who will usually use data one year beyond the sample, however, may wish to account for the correlation in the first and perhaps the second months of the validation period.)

#### **THE UPDATED FIRST TERM ETS LOSS MODEL**

The updated version of the first term ETS loss model is presented in Table A.11. It differs from the original specification in seven ways:

1. The updated version adds a time trend from July 1976 to May 1983.
2. The extended sample deletes six-year enlistees who entered the service before July 1972.
3. The extended sample adds airmen who reenlisted more than 12 months before their originally scheduled ETS.
4. The updated version deletes WBONC as an explanatory variable.
5. The updated version deletes the dummy variable for the operational manning policy period.
6. The updated version deletes the Regular Reenlistment Bonus variable for six-year enlistees.
7. The updated model measures unemployment and the military/civilian wage ratio as of 12 months before the airman's originally scheduled ETS rather than at the beginning of the year at risk in which the airman's first term ends.

The first two columns of Table A.11 contain the coefficients of the updated specification.

The inclusion of the time trend in the specification requires a decision on how the time trend is to be treated in making forecasts for the future. Fortunately, the data dictate

Table A.11

UPDATED SPECIFICATION OF FIRST TERM ETS LOSS MODEL  
(Using YAR 3.0 data for the periods July 1976–May 1985  
and July 1976–June 1983)

Predictor Variables	Variable Name	7607–8505		7607–8306	
		Coeff.	t-stat.	Coeff.	t-stat.
Male	BMALE	0.115	(19.91)	0.130	( 19.43)
Married	MAR	-0.030	(-4.10)	-0.034	(-4.08)
Male and married	MAR*BMALE	-0.088	(-11.13)	-0.080	(-8.78)
Black	BBLK	-0.162	(-41.04)	-0.169	(-37.70)
Log(mil wages/civ wages)	LTMCPAY	-0.750	(-7.63)	-0.779	(-7.75)
4-year $\times$ log(moving average of unemployment)	T4*LTMAU	-0.106	(-7.16)	-0.143	(-7.97)
Six-year enlistee	T6	0.142	(4.23)	0.220	( 5.40)
Half bonus	BH	-0.024	(-2.67)	-0.010	(-0.56)
Bonus level=1	B1	-0.033	(-6.40)	-0.026	(-4.57)
Bonus level>1	BG1	-0.019	(-3.54)	-0.020	(-3.42)
Period of Regular Reenlistment Bonus	BRRB*T4	-0.037	(-4.35)	-0.024	(-2.66)
$\times$ 4-year enlistee					
Time trend through 8305	TREND83	-0.0028	(-15.66)	-0.002	(-12.98)
Time trend from 8305–8505	TTREND2	0.00005	(0.84)	—	—
Mean rate		.451		.473	
Sample size		127157		100301	

a simple treatment; there is no evidence that the time trend continues past June 1983. This is tested for in the model by specifying (1) a trend variable through June 1983 that holds constant at its June 1983 value in subsequent time periods and (2) a trend variable from July 1983 through the end of the sample (before July 1983 this variable's value is zero). The estimated coefficient for the trend from July 1983 onward is very small and has a small standard error. (The estimated annual trend from July 1976 onward is .034. The estimated annual effect after June 1983 is more than three estimated standard deviations below .003.) Therefore the cumulative effect of the time trend through June 1983 is included as a component of the constant term in the forecasting model but no subsequent time trend.

The last two columns of Table A.11 present the parameter estimates for the updated specification over the original sample period (using the the extended sample less the last 23 months of data). Few coefficients differ much between the two data sets. The only appreciable changes are in the unemployment coefficient, which is smaller in the full sample, and in the half-level and one-level bonus variables, which are larger in the full sample. (The change in the six-year enlistee coefficient is largely a reflection of the change in the intercept caused by the change in the unemployment coefficient, which applies only to four-year enlistees.) Nearly all coefficients, especially the half bonus coefficient, are measured with greater estimated precision in the full data set. (The time series analysis of the previous section suggests that the serial correlation in the time series component of the data causes OLS to underestimate by 10 to 20 percent the standard errors of the unemployment and pay coefficients. The other coefficients' estimated standard errors are less affected because they rely more heavily on the cross-sectional component of the data.)

The cross-bonus opportunity variable, WBONC, is deleted from the model because its estimated coefficient is neither large ( $-.01$ ) nor statistically significant. The dummy variable for the operational manning policy period, OPMAN, and the dummy variable for the Regular Reenlistment Bonus Period for six-year enlistees, BRRB, are deleted for the same reason. None of these deletions much affect other coefficients.

The change in when pay and unemployment are measured does affect other coefficients, however. The original specification used the military/civilian wage ratio and unemployment as measured at the beginning of the year at risk in which airmen ended their first term. As a consequence, airmen who left the service or reenlisted more than 12 months before their ETS had different values for the pay and unemployment values than did other airmen with the same originally scheduled ETS who ended their term within 12 months of their originally scheduled ETS. But the early outs and early reenlistments are almost entirely determined by Air Force policy, not by the mean loss rate of an original ETS cohort (a cohort of airmen with the same originally scheduled ETS). Any correlation between the loss rates and pay or unemployment variables for airmen in a single cohort is therefore spurious and should not be allowed to influence the estimates of the coefficients of the pay and unemployment variables. A more satisfactory way of treating pay and unemployment is to use the same measure for all members of a single original ETS cohort. The measures used in the updated specification are the pay

and unemployment variables as measured 12 months before the originally scheduled ETS. This change in the specification causes a 25 percent rise in the coefficient of the unemployment variable, from  $-.085$  to  $-.106$ .

A change not made in the updated specification is to replace ultimate AFSC with some other measure of occupation. In the expanded sample, with the revised specification, an altered occupational specification produced little change in the coefficients of the model. This is only a happenstance, so the design of more satisfactory occupational codes should remain on the agenda for the future.

## **Appendix B**

### **ABBREVIATIONS FOR THE EQUATIONS**

1att2	first term attrition in first two months of term
1att10	first term attrition in remainder of first year of term
1atts	first term attrition in remainder of period preceding the year before the original expiration of term of service (OETS)
1ets	first losses in year preceding OETS
1egs	first term extend given stay
1xlnd	first term extension losses for nondecisionmakers
1xld	first term extension losses for decisionmakers
2att	second term attrition
2ets	second term losses in year preceding OETS
2egs	second term extend-given-stay
2xlnd	second term extension losses for nondecisionmakers
catt	career attrition
cets	career losses in year preceding OETS
cxlnd	career extension losses for nondecisionmakers
cxld	career extension losses for decisionmakers
ret	retirements

## REFERENCES

- Abrahamse, Allan, *Middle-Term Disaggregate Loss Model Test and Evaluation: Description and Results*, The RAND Corporation, N-2688-AF, May 1988.
- Arguden, R. Yilmaz, *Personnel Management in the Military: Effects of Retirement Policies on the Retention of Personnel*, The RAND Corporation, R-3343-AF, January 1986.
- Brauner, Marygail, Michael Murray, Warren Walker, and Elizabeth Davidson, *What's on The Enriched Airman Gain/Loss File*, The RAND Corporation, N-2610-AF, March 1989.
- Carter, Grace, Michael Murray, R. Yilmaz Arguden, Marygail Brauner, Harvey Greenberg, Deborah Skoller, Allan Abrahamse, *Middle-Term Loss Prediction Models for the Air Force's Enlisted Force Management System: Specification and Estimation*, The RAND Corporation, R-3482-AF, December 1987.
- Gotz, Glenn and John J. McCall, *A Dynamic Retention Model for Air Force Officers*, The RAND Corporation, R-3028-AF, December 1984.
- Murray, Michael, Grace Carter, Daniel Relles, Marygail Brauner, Leola Cutler, Deborah Skoller, Warren Walker, *What's on The Year-At-Risk File*, The RAND Corporation, N-2744-AF, March 1989.
- Walker, Warren, and Dan McGary, *Supplemental Historical Data Files for the Enlisted Force Management Project*, The RAND Corporation, N-2844-AF, March 1989.